GMD Research Series

GMD –
Forschungszentrum
Informationstechnik
GmbH

Ralf Thiele

Algorithmen und Bewertungssysteme für die ähnlichkeitsbasierte Proteinstrukturvorhersage

N° 17/1998

GMD – Forschungszentrum Informationstechnik GmbH
Schloß Birlinghoven
D-53754 Sankt Augustin
Germany
Telefon +49 -2241 -14 -0
Telefax +49 -2241 -14 -2618
http://www.gmd.de

Anschrift des Verfassers/Address of the author:
Ralf Thiele
Ebbinghauser Straße 13
D-33178 Borchen
E-mail: RALF.THIELE.RT@bayer-ag.de

Die vorliegende Veröffentlichung entstand im/
The present publication was prepared within:
Institut für Algorithmen und Wissenschaftliches Rechnen (SCAI)
Institute for Algorithms and Scientific Computing
http://www.scai.gmd.de

D5
Die Deutsche Bibliothek - CIP-Kurztitelaufnahme:

Thiele, Ralf:
Algorithmen und Bewertungssysteme für die ähnlichkeitsbasierte Proteinstrukturvorhersage / Ralf Thiele. GMD – Forschungszentrum Informationstechnik GmbH. - Sankt Augustin :
GMD – Forschungszentrum Informationstechnik, 1998
(GMD Research Series ; 1998, No. 17)
Zugl.: Bonn, Univ., Diss., 1998
ISBN 3-88457-341-1

ISSN 1435-2699
ISBN 3-88457-341-1

Abstract

This thesis presents a divide-and-conquer method for protein threading called *recursive dynamic programming* (RDP). Protein threading is one of the most successful methods to detect distant relationships between protein sequences and proteins whose three dimensional structures have been experimentally solved. Such relationships are used to predict structural models for protein sequences. The quality of the derived structural model is mainly determined by the correctness of the mapping of the sequence onto the template structure. Therefore, by calculating high quality sequence structure alignments the RDP method aims at the improvement of the reliability of fold recognition and corresponding model structures. The RDP method works as follows: The protein sequence is mapped onto a potential template structure in a stepwise fashion, similarly to computing local alignments but utilizing different cost functions. RDP, recursively, modifies the template structure in order to account for the mapped residues and searches for significant similarities between the yet unmapped parts of the sequence and the modified template. This recursive process is continued until no significant similarities between the remaining parts of sequence and template are found according to the scoring system in the context of the already mapped parts.
We validate our method on different sets of protein pairs where both structures are known as well as with blind predictions. On standard test sets the RDP method shows significant improvements of the alignment quality in comparison with available state-of-the-art threading tools. As a result of the improved alignment quality the number of distant structural relationships reliably identified in fold recognition experiments is also significantly increased.

Key words: protein structure prediction, fold recognition, sequence-structure alignment, discrete optimization, threading, divide & conquer algorithm, potentials of mean force

Zusammenfassung

In dieser Arbeit wird eine *divide & conquer* Methode für das Sequenzstrukturalignment von Proteinen vorgestellt (*rekursiven dynamischen Programmierung* (RDP)). Sequenzstrukturalignment ist eine der erfolgreichsten Methoden, um entfernte Verwandtschaften zwischen Proteinsequenzen und anderen experimentell strukturaufgeklärten Proteinen zu erkennen, die der Ableitung von Strukturmodellen für die betrachtete Sequenz dienen. Die Qualität der so erhaltenen Strukturmodelle wird wesentlich durch die Korrektheit der Abbildung der Sequenz auf die Strukturen festgelegt, die als Vorlage dienen. Daher zielt die RDP–Methode auf die Berechnung von sehr guten Sequenzstrukturalignments ab, um so die Zuverlässigkeit sowohl der Faltungserkennung als auch der zugehörigen Modellstrukturen zu erhöhen.
Die RDP–Methode bildet die Proteinsequenz schrittweise auf eine bekannte Rückgratstruktur ab, indem sie lokale Alignments mit unterschiedlichen Kostenfunktionen berechnet. RDP modifiziert jeweils die Vorlagestruktur gemäß der bereits abgebildeten Aminosäureresten und sucht dann rekursiv nach signifikanten Ähnlichkeiten zwischen bisher nicht berücksichtigten Teilen in Sequenz und Vorlagestruktur. Dieser rekursive Prozeß wird solange fortgesetzt, bis auch im Kontext der bereits abgebildeten Teile keine signifikanten Ähnlichkeiten zwischen den verbleibenden Teilen von Sequenz und Struktur mehr gefunden werden.
Die RDP–Methode wird sowohl an Beispielen mit bekannten Strukturen als auch an echten Blindvorhersagen validiert. Auf Standardtestmengen liefert die RDP–Methode im Vergleich zu anderen *state–of–the–art*–Sequenzstrukturalignmentverfahren eine signifikante Verbesserung der Alignmentqualität und damit in direkter Konsequenz eine Erhöhung der zuverlässig in Faltungserkennungsexperimenten erkannten entfernten strukturellen Verwandtschaften.

Schlagworte: Proteinstrukturvorhersage, Faltungserkennung, Sequenzstrukturalignment, diskrete Optimierung, Threading, Divide & Conquer–Algorithmus, statistische Potentiale

Vorwort

Die in dieser Arbeit zusammengefaßten Ergebnisse sind während meiner Tätigkeit als wissenschaftlicher Angestellter am Institut für Algorithmen und wissenschaftliches Rechnen der GMD – Forschungszentrum Informationstechnik GmbH in den vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie geförderten Verbundprojekten *Proteine: Sequenz, Struktur und Evolution* (`PROTAL`, FKZ 01 IB 301 A1) und *Target–Identifizierung: Methoden zum Auffinden von Zielproteinen in Genomdaten* (`TargId`, FKZ 0311615) entstanden.

Ich möchte mich an dieser Stelle bei all jenen bedanken, die zum Gelingen der Arbeit beigetragen haben.

Mein besonderer Dank gilt meinem Doktorvater Thomas Lengauer, der mir die Möglichkeit zur Forschungsarbeit in seiner Arbeitsgruppe gab und bei der Erstellung der Arbeit die notwendige Entfaltungsfreiheit ließ, mir aber stets als Ratgeber und motivierender Diskussionspartner zur Seite stand.

Mein ganz besonderer Dank gilt meinem Kollegen Ralf Zimmer, der durch seine konstruktiv kritische Betrachtung der geleisteten Forschungsarbeit und gleichzeitig motivierende Art im entscheidenden Maße zum Gelingen dieser Arbeit beigetragen hat.

Mein Dank gilt auch Heinz–Theodor Mevissen, der grundlegend an der Implementierung der `ToPLign`–Umgebung beteiligt war, die in vielen Teilen die Arbeitsgrundlage für die Entwicklung und Implementierung der von mir entwickelten RDP–Methode war.

Sankt Augustin, im Mai 1998 Ralf Thiele

Inhaltsverzeichnis

Kapitel 1
Einleitung

Proteine spielen eine grundlegende Rolle beim Ablauf der meisten biologischen Prozesse und erfüllen bei der Steuerung der komplexen Abläufe biochemischer Reaktionen mannigfaltige Funktionen. Proteine sind daher *die* Zielmoleküle (*targets*) für nahezu alle heute eingesetzten Arzneimittel. Das für die Entwicklung von neuartigen biologischen Wirkstoffen aber auch für den wissenschaftlichen Fortschritt wesentliche Verständnis ihrer Funktion kann letztlich nur auf Basis der Kenntnis ihrer räumlichen Struktur gewonnen werden. Die experimentelle Aufklärung der dreidimensionalen Struktur von Proteinen ist jedoch sehr aufwendig und kann daher nicht mit der Aufklärung der Sequenzen der Proteine Schritt halten, die gegenwärtig unter anderem in den zahlreichen Genomprojekten erfolgt.
Diese Arbeit befaßt sich mit Algorithmen für ein zentrales Problem der molekularen Bioinformatik [199, 201], der theoretischen Proteinstrukturvorhersage: der Vorhersage der dreidimensionalen Struktur eines Proteins aus seiner Aminosäuresequenz. Der entwickelte Lösungsansatz basiert methodisch auf der ähnlichkeitsbasierten Proteinstrukturvorhersage, die gegenwärtig den einzigen in vielen Fällen erfolgreichen Weg zur Vorhersage der dreidimensionalen Struktur von Proteinen darstellt.
Die theoretische Proteinstrukturvorhersage und insbesondere die ähnlichkeitsbasierte Strukturvorhersage stehen gegenwärtig im besonderem Interesse der internationalen Forschung. Dies äußert sich sowohl in der großen Anzahl relevanter Publikationen (siehe Kapitel 4 und 5) als auch in der Existenz des internationalen Proteinstrukturvorhersagewettbewerbs „*Critical Assessment of Methods of Protein Structure Prediction*“ [239, 240].
Der Schwerpunkt dieser Arbeit liegt auf der Entwicklung von Algorithmen und Bewertungssystemen zur Abbildung von Sequenzen auf experimentell aufgeklärte Proteinstrukturen, allgemein als Sequenzstrukturalignment oder auch *Threading* bezeichnet. Die Berechnung dieser Abbildung stellt den ersten und damit grundlegenden Arbeitsschritt der ähnlichkeitsbasierten Proteinstrukturvorhersage dar. Fehler, die hier gemacht werden, können in den nachfolgenden Schritten nur schwer und sehr häufig nicht mehr behoben werden.
Die in dieser Arbeit entwickelte Methode der Rekursiven Dynamischen Programmierung (RDP) zielt in erster Linie auf die Berechnung strukturrichtiger Sequenzstrukturalignments, die sowohl zur Vorhersage von strukturellen Verwandtschaften als auch als Startpunkt für die vergleichende Modellierung dienen. Das algorithmische Grundgerüst der RDP–Methode ist ein modular aufgebauter *divide & conquer*–Algorithmus, der auch für andere bioinformatische Vergleichsprobleme, wie das Proteinstrukturalignment und das multiple Sequenzalignment, angewen-

det werden kann.
Bei der Berechnung strukturrichtiger Alignments kommen unterschiedliche Bewertungssysteme zum Einsatz (siehe Kapitel 5), angefangen von einfachen Aminosäureaustauschmatrizen bis hin zu komplexen Pseudoenergiepotentialen, die Wechselwirkungen zwischen zwei oder mehr Aminosäureresten in einer dreidimensionalen Struktur bewerten und aufgrund ihrer Nichtlokalität das Sequenzstrukturalignmentproblem zu einem *NP*–vollständigen Problem machen [193].
Im Unterschied zu anderen, vom Grundkonzept ähnlichen Verfahren [195, 369], die das fragmentbasierte Sequenzstrukturalignmentproblem adressieren (siehe Abschnitt 4.2.1), soll die RDP–Methode [340] den *divide & conquer*–Ansatz zur Lösung des allgemeinen, nicht fragmentbasierten Sequenzstrukturalignmentproblems eingesetzt werden.
Die fragmentbasierte Modellierung des Sequenzstrukturalignmentproblems bietet aus berechnungstechnischer Sicht den Vorteil, daß in *Branch&Bound*–Verfahren aufgrund von in diesem Spezialfall ableitbaren Schranken große Teile des zu durchsuchenden Lösungsraums nicht enummeriert werden müssen. Durch diese Schranken können häufig mit einem heuristischen Ansatz sogar optimale Lösungen in akzeptabler Zeit berechnet werden [195].
Die Reduktion des Sequenzstrukturalignmentproblems auf die Berechnung einer Abbildung von durch die Sekundärstrukturelemente definierten Fragmenten auf Sequenzpositionen stellt jedoch eine starke Vereinfachung der biologischen Realität dar, da weder die Struktur noch die Funktion eines Proteins allein durch seine Sekundärstrukturelemente festgelegt werden.
Die RDP–Methode soll daher eine präzisere, alle Bereiche eines Proteins umfassende und damit biologisch relevantere Problemmodellierung verwenden und kann daher nicht alle berechnungstechnischen Vorteile des fragmentbasierten Ansatzes nutzen.
Mit der RDP–Methode soll eine Heuristik entwickelt werden, die weniger auf die Optimierung einer globalen Bewertungsfunktion, sondern vielmehr auf die Berechnung eines biologisch sinnvollen Strukturmodells abzielt, zu dessen Berechnung einerseits alle zwischen der Sequenz und der Struktur vorhandenen Ähnlichkeiten genutzt werden sollen, das aber andererseits auch nur die Eigenschaften der Struktur übernehmen soll, die zwischen der Sequenz und Struktur konserviert sind.
Auf die Garantie der Optimalität der gefundenen Lösung bezüglich des Bewertungssystems kann zugunsten biologischer Randbedingungen verzichtet werden, da allen bisher bei der Lösung des Sequenzstrukturalignmentproblems verwendeten Bewertungssystemen gemeinsam ist, daß sie mit statistischen Methoden aus diskretisierten Beschreibungen der biologischen und biochemischen Realität abgeleitet sind und damit nur eine Approximation der die Proteinfaltung determinierenden Energien sind. Im Unterschied zu vielen anderen Sequenzstrukturalignmentverfahren [48, 168, 195, 321] soll die RDP–Methode daher eine gemischte Bewertungsfunktion verwenden, die neben Anteilen, die die Sequenzähnlichkeit

bewerten, auch das Kontaktkapazitätspotential `CCP` [5] enthält, welches im wesentlichen den Hydrophobizitätsanteil kodiert, der in Wechselwirkungspotentialen unterrepräsentiert ist.
Aufgrund der im Vorfeld dieser Arbeit erkannten Schwächen globaler Bewertungsfunktionen beim Sequenzstrukturalignment soll die RDP–Methode nicht unbedingt nach *optimalen*, sondern nach Lösungen mit hinreichend *guter* Bewertung suchen, die zusätzliche Kriterien und Randbedingungen erfüllen. Zum Beispiel sind die aktiven Zentren zwischen verwandten Proteinen in vielen Fällen in ihrer Aminosäuresequenz wesentlich stärker konserviert als dies für die Gesamtsequenzen der Fall ist. Ein berechnetes Modell sollte dies in jedem Falle wiedergeben. Die Idee, rekursiv signifikante Teillösungen mit sich adaptiv anpassenden Bewertungsfunktionsbestandteilen zu suchen, soll diese biologische Eigenschaft ausnutzen und möglichst auch dann zu aussagekräftigen Modellen des aktiven Zentrums führen, wenn sich die strukturelle Ähnlichkeit auf Teilstrukturbereiche beschränkt.
Bereiche in Sequenz und Struktur, zwischen denen mit keinem der Bewertungsfunktionsbestandteile eine signifikante Ähnlichkeiten nachgewiesen werden kann, sollen von der RDP–Methode auch nicht aliniert werden, so daß sie als sogenannte Insertionen und Deletionen übrigbleiben. Die Idee ist, daß die rekursive Problemlösungsstrategie der RDP–Methode genau dann stoppt, wenn keine signifikanten Ähnlichkeiten mehr gefunden werden.
Dadurch wird das Ziel verfolgt, das Alignment auf die Bereiche einzuschränken, deren Ähnlichkeit sich auch durch eine experimentelle Strukturaufklärung belegen läßt. Damit würde die RDP–Methode einen wesentlichen Fortschritt gegenüber bisherigen Methoden erzielen, da fälschlicherweise durch ein Alignment suggerierte Ähnlichkeiten sich in der Regel in den nachfolgenden Strukturvorhersageschritten entweder als sehr störend erweisen beziehungsweise im schlimmsten Fall sogar zu falschen Modellen führen.
Ein weiteres Ziel ist es, durch diese Vorgehensweise Bestrafungsterme für Insertionen und Deletionen (Gapkosten) bei der Berechnung strukturrichtiger Sequenzstrukturalignments überflüssig zu machen oder zumindest ihren Einfluß auf die Ergebnisse drastisch zu reduzieren. Eine derartige Eigenschaft der RDP–Methode wäre von besonderem Vorteil, da die geeignete Wahl von Gapkosten in den bisher vorgeschlagenen Methoden – sei es zum Sequenzalignment oder Sequenzstrukturalignment – problematisch ist [380]. Außerdem könnten dann die Bestrafungsterme für Insertionen und Deletionen als orthogonales Kriterium bei der Erkennung von verwandten Proteinen eingesetzt und so die Erkennungsrate weiter verbessert werden.
Außerdem soll mit der RDP–Methode ein Verfahren entwickelt werden, das durch die explizite Einbeziehung von Randbedingungen für die Modellierbarkeit von Schleifen in den Berechnungsprozeß garantiert, daß ein berechnetes Sequenzstrukturalignment immer in ein zulässiges Strukturmodell mit geschlossenem Rückgrat übersetzt werden kann.

Mit der RDP–Methode werden folgende Fortschritte in der theoretischen Proteinstrukturvorhersage angestrebt:

- Die Genauigkeit der als Grundlage für die vergleichende Modellierung von Proteinstrukturen dienenden Abbildungen der zu modellierenden Sequenz auf die Modellstruktur, soll durch die RDP–Methode signifikant verbessert werden.
- Die Anwendbarkeit der vergleichende Modellierung soll durch die RDP–Methode auf Proteinsequenzen ausgedehnt werden, zu denen bisher keine strukturaufgeklärten Proteine mit signifikanter Sequenzähnlichkeit bekannt sind.
- Von der besseren Alignmentqualität der RDP–Methode sollte in den Fällen nicht auf Sequenzebene nachweisbarer Ähnlichkeiten auch die Erkennung der zu einer Sequenz ähnlichen Faltungen in der Strukturdatenbank profitieren.

Wie die in Kapitel 8 diskutierten Ergebnisse zeigen, ist die RDP–Methode allen diesen Zielen gerecht geworden. Experimente an bekannten Strukturen zeigen, daß die Alignmentqualität durch die RDP–Methode für Beispiele mit Sequenzidentitäten bis 80% (über 80% ist die Berechnung des Alignments trivial) wesentlich verbessert werden kann. Im für die bisherigen Methoden schwierigen Bereich unter 30% Sequenzidentität liefert die RDP–Methode für repräsentative Testmengen doppelt soviele gute Alignments wie bisher verwendete Methoden und erweitert damit wesentlich den Anwendungsbereich der vergleichenden Modellierung.
Die mit der entwickelten Methode in der Modellqualität und der Faltungserkennung erzielten Fortschritte werden nicht nur anhand von Datenbankexperimenten, sondern auch in echten Blindvorhersagen dokumentiert, in denen sich die RDP–Methode insbesondere durch ein besonders gutes Modell für die aktiven Zentren der vorhergesagten Proteinstrukturen hervorhebt.

Die Arbeit gliedert sich wie folgt:

Kapitel 2 stellt den Bezug der in dieser Arbeit vorgestellten theoretischen Strukturvorhersagemethoden zur Biologie der Gene und Proteine und zu den experimentellen Methoden her. Dabei beschreibt Abschnitt 2.1 den chemischen Aufbau und die dreidimensionale Struktur von Proteinen. Im Abschnitt 2.2 werden die experimentellen Möglichkeiten zur Proteinstrukturbestimmung, deren Grenzen und die damit einhergehende Notwendigkeit theoretischer Vorhersagemethoden thematisiert. Abschnitt 2.3 verdeutlicht, daß theoretische Vorhersagemethoden zunächst auf die Vorhersage des Endergebnisses des Faltungsprozesses zielen sollten, da der eigentliche Faltungsvorgang nur schwer experimentell beobachtet werden kann und aufgrund seiner Komplexität bisher nicht durch eine allgemein akzeptierte Theorie erklärt werden kann.

Kapitel 3 begründet die Relevanz des Strukturvorhersageproblems durch die Lücke zwischen den bekannten Sequenzen und aufgeklärten Proteinstrukturen. Desweiteren wird in Abschnitt 3.2 mit einer ersten formalen Problemdefinition versucht, aufzuzeigen, warum die Optimierung äußerst komplexer Energiefunktionen mit Simulationsmethoden kein gangbarer Weg zur Lösung des Strukturvorhersageproblems ist. Abschnitt 3.3 belegt an strukturellen Vergleichen und Klassifizierungen, daß die ähnlichkeitsbasierte Proteinstrukturvorhersage einen gangbaren und den derzeit einzigen erfolgversprechenden Lösungsweg darstellt.
In Kapitel 4 werden verschiedene, heute zur Proteinstrukturvorhersage eingesetzte Verfahren diskutiert. Von diesen sind die vergleichende Modellierung (Abschnitt 4.1) und die Sekundärstrukturvorhersage (Abschnitt 4.3.1) heute als *state–of–the–art* anzusehen. In Abschnitt 4.3 wird motiviert, warum *ab initio*–Tertiärstrukturvorhersagemethoden bislang keinen Lösungweg darstellen. Die vergleichende Modellierung im klassischen Sinn kann nur angewendet werden, wenn signifikante Sequenzähnlichkeiten zu experimentell aufgeklärten Strukturen vorhanden sind. Ziel der in Abschnitt 4.2 beschriebenen Faltungserkennungsverfahren ist es, strukturelle Verwandtschaften *ohne* signifikante Sequenzähnlichkeiten aufzudecken. Die in dieser Arbeit vorgestellte RDP–Methode hat neben der Erkennung derartiger Verwandtschaften auch die Optimierung der für die Modellbildung verwendeten Alignments zum Ziel. Daher kann sie sowohl zur Faltungserkennung als auch zur Verbesserung der typischerweise auf Sequenzalignments (siehe Abschnitt 4.1.2) aufbauenden vergleichenden Modellierung eingesetzt werden.
Die in der RDP–Methode eingesetzten Bewertungssysteme werden in Kapitel 5 vorgestellt. In Abschnitt 5.1 werden die sequenzabhängigen Bewertungssysteme untersucht. Die Grundlagen für die statistische Ableitung von Ein– beziehungsweise Mehrkörperpotentialen werden in Abschnitt 5.2 dargestellt. Die Abschnitte 5.2.3 und 5.2.4 stellen die adaptierten und teilweise neuentwickelten Potentiale vor, die in der RDP–Methode eingesetzt werden.
In Kapitel 6 wird die neue RDP–Methode algorithmisch motiviert und ihre Ablaufstruktur auf hohem Abstraktionsniveau vorgestellt. Das algorithmische Prinzip der RDP–Methode kann zur Lösung verschiedener *NP*–vollständiger Vergleichsprobleme der molekularen Bioinformatik eingesetzt werden.
Kapitel 7 beschreibt die Anpassung der RDP–Methode auf das in dieser Arbeit zentrale Anwendungsproblem, das Sequenzstrukturalignment. In Abschnitt 7.1 werden die verwendeten Datenstrukturen vorgestellt. Die Bestimmung sinnvoller Teillösungen für Teilprobleme, die während der rekursiven Problemlösung entstehen, ist Gegenstand von Abschnitt 7.2. In Abschnitt 7.3 werden Wege aufgezeigt, wie der Suchraum durch Filterprozeduren eingeschränkt werden kann, die unter anderem biologische Randbedingungen einbringen. Die Abschnitte 7.4 und 7.5 befassen sich mit der Aufteilung in Unterprobleme und der Festlegung der Abarbeitungsreihenfolge der Unterprobleme. Kapitel 7 schließt mit einer Beschreibung der Zusammenfassungsphase (Abschnitt 7.6) der RDP–Methode und einem er-

sten Versuch der Kalibrierung der Gewichtung der Kostenfunktionsbestandteile zur Berechnung strukturrichtiger Sequenzstrukturalignments (Abschnitt 7.7).
Kapitel 8 belegt den Fortschritt, der mit der RDP–Methode für die Proteinstrukturvorhersage erreicht wird, anhand von typischen Anwendungsszenarien und repräsentativen Beispielmengen. Nach der Einführung verschiedener Erfolgskriterien (Abschnitt 8.1) für Proteinstrukturvorhersagemethoden und der Beschreibung der generellen Vorgehensweise bei einer Blindvorhersage (Abschnitt 8.2) wird in Abschnitt 8.3 die Qualität der mit der RDP–Methode berechneten Sequenzstrukturalignments anhand von Beispielen, wo auch die Struktur des zweiten Proteins bekannt ist, untersucht und mit Ergebnissen anderer Methoden verglichen. Abschnitt 8.4 ist der Faltungserkennung gewidmet. Die Abschnitte 8.5 und 8.6 beschreiben echte Blindvorhersagen, für die die RDP–Methode eingesetzt wurde.
Die Arbeit schließt mit einem Ausblick auf zukünftige Erweiterungen und Verbesserungen in Kapitel 9 und einer Zusammenfassung in Kapitel 10.

Kapitel 2

Proteine: Chemischer Aufbau und dreidimensionale Struktur

2.1 Proteine und ihre strukturbeschreibenden Elemente

Beim Ablauf aller biologischen Prozesse spielen Proteine eine grundlegende Rolle. Da die biologischen Prozesse in ihrer Gesamtheit *das Leben* ausmachen, werden Proteine vielfältig auch als die Grundbausteine des Lebens bezeichnet. Sie erfüllen bei der Steuerung der komplexen Abläufe biochemischer Reaktionen mannigfaltige Funktionen. Proteine

- sind *enzymatische Katalysatoren* für nahezu alle chemischen Reaktionen in Organismen,
- sind *Botenstoffe* und *Rezeptoren* in der Übertragung von Signalen (Reize werden von Nervenzellen über Rezeptorproteine aufgenommen),
- dienen als *Speicher*– und *Transportmedium* biologisch relevanter Substanzen (zum Beispiel Sauerstofftransport durch Hämoglobin),
- sind passive aber dehnbare Bauelemente von Sehnen, Bändern und Muskeln,
- dienen als Strukturproteine (zum Beispiel Kollagen in Binde– und Stützgeweben),
- bilden als hochspezifische *Antikörper* die Immunabwehr von Organismen aus und
- kontrollieren das Zellwachstum und die Zelldifferenzierung [36, 14, 353].

Proteine sind linear aufgebaute Ketten von Aminosäureresten, die durch kovalente Bindungen, die als *Peptidbindungen* bezeichnet werden, verknüpft sind. Die Länge dieser Ketten variiert von einigen wenigen Resten (kurzen Peptiden) bis zu einigen tausend Resten.
Eine Aminosäure besteht aus der Aminogruppe (NH_2), dem C_α–Atom mit einem Wasserstoffatom, der Carboxylgruppe ($COOH$) und der sogenannten Seitenkette. Von den aus der belebten Natur bisher bekannten 260 Aminosäuren werden in allen bekannten Lebensformen in der Regel nur 20 durch Tripel aus Nukleinsäuren in den Genen kodiert [14, 353]. Diese 20 Aminosäuren (siehe Tabelle 2.1 werden daher auch als *proteinogene* Aminosäuren bezeichnet. Bei einigen Pro– und Eukaryonten findet sich zusätzlich Transferribonukleinsäuren, die für die seltene Aminosäure Selenocystein kodieren [353], bei der das Schwefelatom des Cysteins durch ein Selenatom ersetzt ist. Darüber hinaus finden sich in Proteinen noch

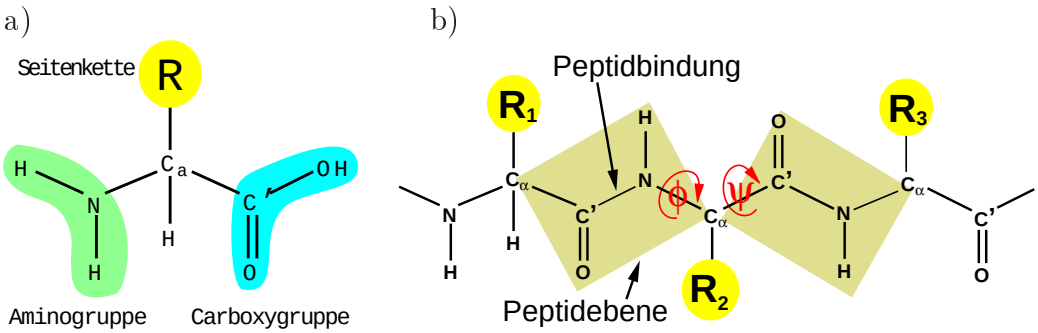


Abbildung 2.1: a) Schematischer Aufbau von Aminosäuren; b) Verknüpfung einzelner Aminosäuren mittels sogenannter Peptidbindung: Die Carboxylgruppe und die Aminogruppe des nächsten Restes bilden die sogenannte Peptidebene (schattiert dargestellt). Die Bindungen $C_\alpha N$ und $C_\alpha C'$ sind frei drehbar, die Winkel werden als ψ– und ϕ–Winkel bezeichnet.

weitere Aminosäuren, die aber in allen bekannten Fällen Produkte spezifischer posttranslationaler Modifikationen der proteinogenen Aminosäuren sind (siehe dazu auch [353]).
Die proteinogenen Aminosäuren unterscheiden sich bis auf die Aminosäure Prolin nur in ihren Seitenketten. Prolin enthält als einzige der proteinogenen Aminosäuren eine sekundäre, zyklische Aminogruppe. Die Aminosäure Glycin nimmt eine Sonderrolle ein, da ihre Seitenkette nur aus einem Wasserstoffatom besteht. Abbildung 2.1 zeigt den schematischen Aufbau einer einzelnen Aminosäure (Teil a)) und einen Ausschnitt aus einer durch Peptidbindungen verknüpften Kette von Aminosäuren (Teil b)).

hydrophobe Aminosäuren			geladene Aminosäuren			polare Aminosäuren		
Alanin	Ala	A	Arginin	Arg	R	Asparagin	Asp	N
Isoleucin	Ile	I	Aspartat	Asp	D	Cystein	Cys	C
Leucin	Leu	L	Glutamat	Glu	E	Glutamin	Gln	Q
Methionin	Met	M	Lysin	Lys	K	Histidin	His	H
Phenylalanin	Phe	F				Serin	Ser	S
Prolin	Pro	P				Threonin	Thr	T
Valin	Val	V				Tryptophan	Trp	W
						Tyrosin	Tyr	Y
Glycin	Gly	G						

Tabelle 2.1: Die 20 proteinogenen Aminosäuren und ihre Drei– und Einbuchstabencodes.

Die 20 proteinogenen Aminosäuren sind L–Aminosäuren (das heißt Aminosäuren, die am α–Kohlenstoffatom ein asymmetrisches Zentrum mit L–Konfiguration

besitzen [14]) und bestehen aus den Elementen Kohlenstoff (C), Wasserstoff (H), Sauerstoff (O) und Stickstoff (N). Das Element Schwefel kommt nur in den Aminosäuren Cystein und Methionin vor. Die Seitenketten unterscheiden sich in Größe, Gestalt, Ladung und chemischer Reaktivität. Tabelle 2.1 zeigt die 20 Aminosäuren mit ihren zugehörigen Drei– und Einbuchstabencodes. Die Einteilung orientiert sich an den Eigenschaften Hydrophobizität, Polarität und polaren Teilladungen. Abbildung 2.2 zeigt den chemischen Aufbau der proteinogenen Aminosäuren.

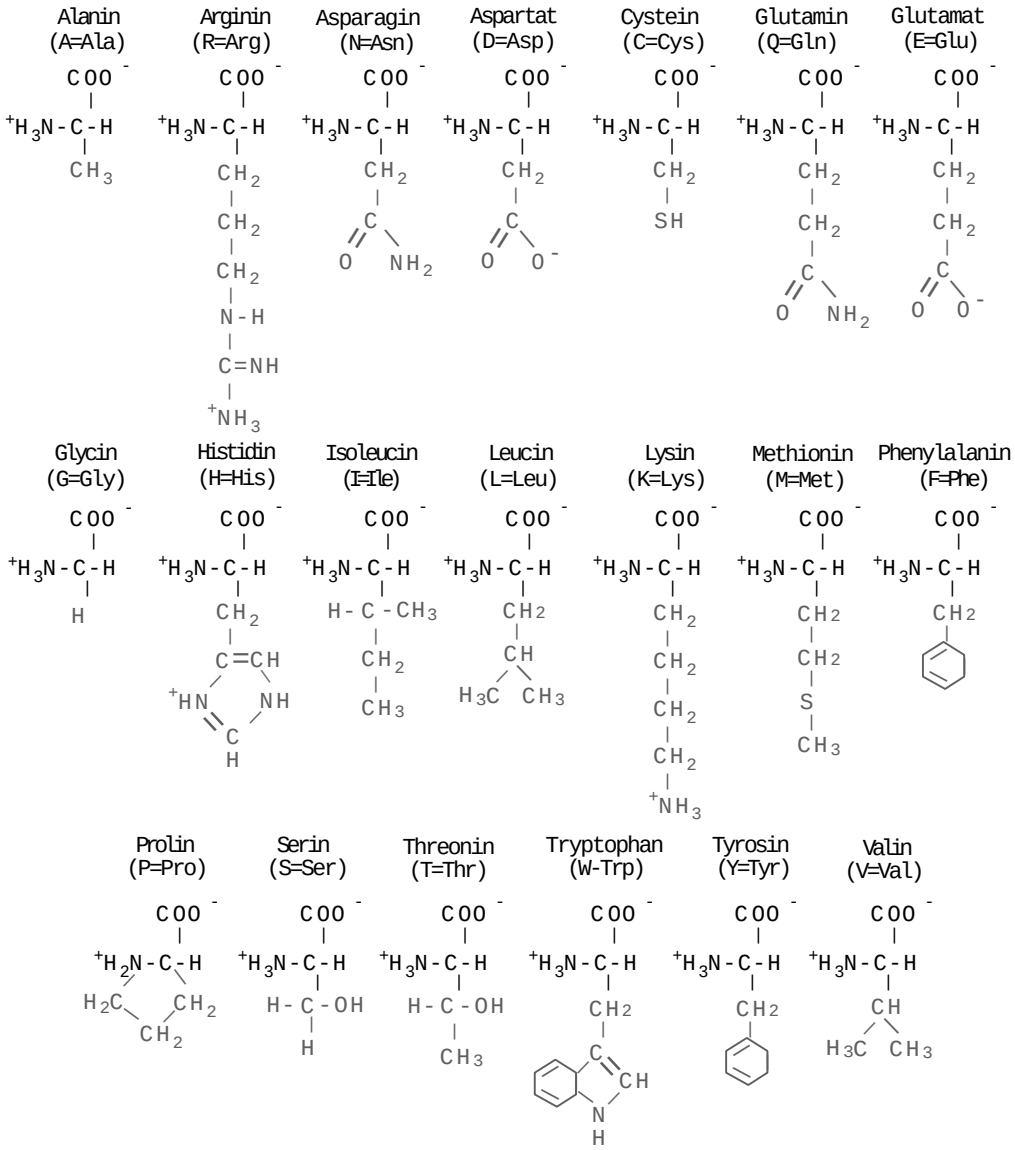


Abbildung 2.2: Die 20 proteinogenen Aminosäuren bezeichnet durch ihren Namen (zugehöriger Drei– und Einbuchstabencode in Klammern). Die Hauptkettenatome sind in schwarz, Seitenkettenatome in grau dargestellt.

Die Beschreibung von Proteinen erfolgt – wie die Beschreibung von linearen Makromolekülen im allgemeinen – auf den folgenden vier Beschreibungsebenen [14]:

1. Die **Primärstruktur** bezeichnet die lineare Abfolge der einzelnen Monomerbausteine. Die Primärstruktur von Proteinen wird durch die *Aminosäuresequenz* beschrieben.

2. **Sekundärstrukturen** sind diejenigen Strukturen von linearen Makromolekülen, die ganz oder zu einem erheblichen Teil durch Wasserstoffbrückenbindungen bedingt sind. Bei Proteinen wird die Sekundärstuktur auch als die lokale räumliche Anordnung der Peptidkette ohne Berücksichtigung der durch die Seitenketten bedingten Interaktionen definiert. Die wesentlichen Sekundärstrukturelemente in Proteinen sind Helix– und Strangstrukturen.

3. Die **Tertiärstruktur** ist die spezifische dreidimensionale Anordnung linear aufgebauter Makromoleküle zu übergeordneten, räumlichen Strukturen.

4. Als die **Quartärstruktur** eines Proteins bezeichnet man seinen Aufbau aus zwei oder mehreren gleichen Untereinheiten (Homomer) oder ungleichen Untereinheiten (Heteromer), die über nicht–kovalente Wechselwirkungen oder Disulfidbindungen miteinander verbunden sind.

Die Funktion eines Proteins kann in der Regel nur anhand seiner dreidimensionalen Struktur im Detail verstanden und erklärt werden. Doch aus der sequentiellen Anordnung der Aminosäurereste und aus der Kenntnis lokaler struktureller Anordnungen können in manchen Fällen schon Aussagen über die mögliche Aufgabe eines bestimmten Proteins abgeleitet werden.
Im folgenden wird erläutert, wie Proteine anhand dieser vier Beschreibungsebenen charakterisiert werden können und welche experimentellen Verfahren zu ihrer Bestimmung angewendet werden. Eine ausführlichere Beschreibung findet sich unter anderem in [40].

2.1.1 Primärstruktur

Die Primärstruktur eines Proteins ist seine Aminosäuresequenz. Die Sequenzen der natürlich vorkommenden Proteine sind in den Genomen der zugehörigen Organismen codiert. Das Genom eines Organismus besteht aus einer Vielzahl von Genen, wobei ein Gen ursprünglich als ein Abschnitt der *Desoxyribonukleinsäure* (DNA) bzw. *Ribonukleinsäure* (RNA) definiert ist, der bestimmte erblich bedingte Strukturen oder Funktionen des Organismus codiert [14]. Neuere Untersuchungen haben jedoch ergeben, daß ein Gen auch auf mehrere Bereiche eines Chromosoms verteilt sein kann.
Der Code, der die Übersetzung der Nukleotidsequenz eines Protein–Genes in die Aminosäuresequenz der Proteine beschreibt, ist bekannt als der *genetische Code*

(siehe Abbildung 2.3). 1961 kamen Brenner und Crick [71] zu der Erkenntnis, daß in diesem Code eine Aminosäure durch ein DNA–Triplet, dem sogenannten *Codon*, codiert ist. Nirenberg und Matthaei [255] fanden 1961, daß ein Codon aus Uracil–Bausteinen die Aminosäure Phenylalanin codiert. Die Aufklärung des genetischen Codes wurde 1966 durch weitere Arbeiten der Gruppe um Nirenberg [254] und durch Arbeiten von Khorana *et al.* [256] vervollständigt. In diesem Code codiert jedes Codon eindeutig eine bestimmte Aminosäure. Die meisten Aminosäuren werden jedoch durch verschiedene Codons codiert, und nicht alle Bereiche eines Genoms codieren auch Proteine. Die proteincodierenden Bereiche sind durch spezielle *Start–* und *Stopcodons* in der Genomsequenz gekennzeichnet.

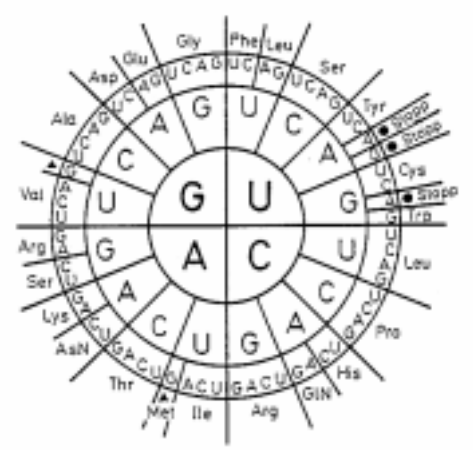


Abbildung 2.3: Der genetische Code: drei aufeinanderfolgende RNA–Nukleotide (*Triplet*) codieren eine Aminosäure.

Bei der Synthese von Proteinen in der Zelle wird zunächst einer der DNA–Stränge in die sogenannte *Messenger*–RNA (kurz auch mRNA) transkribiert (siehe [318] Abbildung I.23 Seite 29 und Seiten 119 ff.). RNA–Polymerasen katalysieren die Synthese der mRNA–Ketten, indem sie die Nukleotidsequenz eines DNA–Strangs durch komplementäre Basenpaarungen kopieren.
Der komplexe Vorgang der *Translation* der mRNA in Proteine findet an den Ribosomen statt. Die Translation beginnt, sobald sich diese an die mRNA angeheftet haben. Bei der schrittweisen Wanderung des Ribosoms, das eine Ansammlung von mehr als 50 Proteinen [318] und drei unterschiedlichen RNA–Typen ist, um ein Codon wird die Proteinkette jeweils um eine Aminosäure verlängert. Die Aminosäuren sind dabei zunächst an die sogenannte *Transfer*–RNA (tRNA) gebunden. Diese Transfer–RNA enthält ein Triplet (*Anticodon*), das komplementär zu dem die entsprechende Aminosäure codierenden Codon in der mRNA ist. Durch diese komplementäre Basenpaarung wird die richtige Aminosäure in die richtige Position gebracht, um an die vom Aminoterminus wachsende Proteinkette

angeknüpft zu werden. Aufgrund dieses gerichteten Wachstums werden Proteinsequenzen im allgemeinen auch als vom Aminoterminus zum Carboxylterminus gerichtete Kette betrachtet.
Dieser Prozeß der Proteinsynthese wird in der modernen Molekularbiologie ausgenutzt, um zum Beispiel für Experimente oder auch zur Strukturaufklärung genügend große Mengen eines bestimmten Proteins herzustellen. Dazu wird der dieses Protein codierende Genabschnitt aus der Zelle, in der das Protein normalerweise exprimiert wird, extrahiert und in das Genom eines Bakteriums (zum Beispiel *Escherichia coli*) eingepflanzt. Unter der Voraussetzung, daß bei der Klonierung und der Expression keine Problem auftreten, produzieren dieses Bakterium und alle seine Nachkommen das gewünschte Protein.
Fortschritte in der Gentechnologie ermöglichen aber nicht nur die Herstellung von Proteinen in ausreichenden Mengen sondern werden auch genutzt, um die Aminosäuresequenz von Proteinen zu bestimmen. Die Sequenzierung von Nukleotidsequenzen ist weitgehend automatisiert [158]. Zum gegenwärtigen Zeitpunkt sind bereits ganze Genome (zum Beispiel das des Eukaryonten Hefe *Saccharomyces cerevisiae* [63, 121], das der Bakterien *Escherichia Coli* [32]), *Helicobacter pylori* [341], *Bacillus Subtilis* [190] und *Borrelia burgdorferi* [103] und das der Archaebakterien *Methanococcus jannaschii* [50] und *Archaeoglobus fulgidus* [180]) sequenziert worden, andere Genome (zum Beispiel das des Ringwurms *Caenorhabditis elegans*) stehen kurz vor der Vervollständigung und auch die Sequenzierung des menschlichen Genoms [130] soll spätestens im Jahr 2005 abgeschlossen sein [120]. Zu diesem Fortschritt hat neben der Weiterentwicklung und Automatisierung der Sequenziertechnik wesentlich die Entdeckung der Polymerasekettenreaktion [241] beigetragen, die eine automatisierte Vervielfältigung von DNA–Sequenzen ermöglicht, die vorher mühsame Laborarbeit bedeutete.
Die Ermittlung der Primärstuktur eines Proteins kann aber auch an gereinigten Proteinen direkt erfolgen. Dazu wird das Protein durch Proteasen oder durch die Bromcyan–Reaktion (selektive Spaltung an Methioninresten) in eine Serie von überlappenden Teilpeptiden gespalten. Die Sequenz dieser Teilpeptide kann dann durch die Technik des Edmanschen Abbaus bestimmt werden, bei dem jeweils die endständige Aminosäure freigesetzt wird. Obwohl diese Technik in Form des Gasphasen–Sequenator weitgehend automatisiert ist, bleibt das Verfahren sehr aufwendig [14] und ist daher heute fast nur von historischer Bedeutung. Auch wenn sich mit der neuentwickelten Methode der Nanoelektrospray–Massenspektrometrie [364] die Sequenz eines Proteins unter Verwendung von nur wenigen Nanogramm Protein ermitteln läßt, wird dennoch heute die Proteinsequenz in der Regel aus der Basensequenz der zugrundeliegenden Gene abgeleitet.

2.1.2 Sekundärstruktur

Die am häufigsten in Proteinen auftretenden Sekundärstrukturelemente sind die rechtsgängige α–*Helix* und der β–*Strang*.

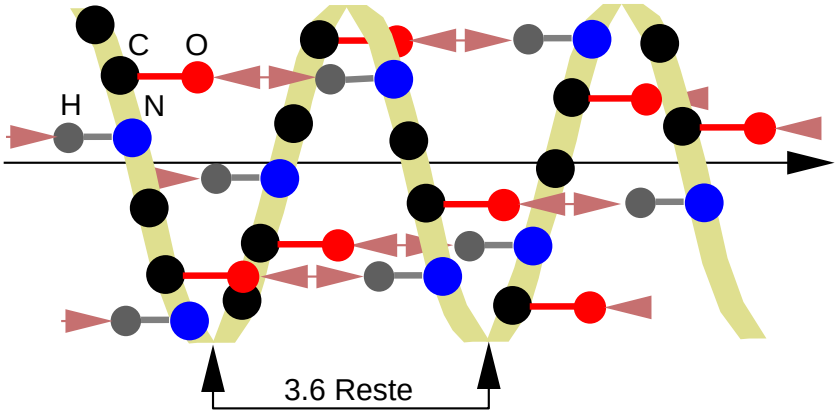


Abbildung 2.4: α–Helix mit Hauptkettenatomen und den für eine Helix typischen Wasserstoffbrückenbindungen.

In der α–Helix bildet das Rückgrat der Peptidkette eine Spirale mit ungefähr 3.6 Aminosäureresten pro Windung. Stabilisiert wird diese Anordnung durch *Wasserstoffbrücken* zwischen der NH–Gruppe der einen Peptidgruppe und der CO–Gruppe der dritt– und/oder viertnächsten Peptidgruppe in der Peptidsequenz, wobei die Seitenketten nach außen ragen. Die seltener vorkommenden 3_{10}– und π–Helices sind durch Wasserstoffbrücken zwischen der einen Peptidgruppe und der dritt– beziehungsweise fünftnächsten Peptidgruppe charakterisiert.
Die Aminosäuren haben unterschiedliche Präferenzen, in α–Helices vorzukommen. Die Aminosäure Prolin unterbricht Helices, da sie zum einen durch ihren Aufbau die freie Drehbarkeit der Peptidkette einschränkt und zum anderen auch kein Wasserstoffatom für die Ausbildung einer Wasserstoffbrücke beisteuern kann. Helices haben typischerweise eine Länge von acht bis dreißig Resten.
Wie Abbildung 2.4 zeigt, haben alle Wasserstoffbrücken in einer α–Helix die gleiche Ausrichtung entlang der Achse der Helix. Die Dipole der einzelnen Peptide, gegeben durch die unterschiedliche Polarität von NH– und CO–Gruppe, verstärken sich so zu einem Gesamtdipol der Helix mit einer positiven Teilladung am Aminoterminus und einer negativen Ladung am Carboxylterminus.
Im Gegensatz zu Helices ist das Rückgrat der Peptidkette in dem zweiten häufigen Sekundärstrukturelement, dem β–*Strang* (Abbildung 2.5), nicht aufgewunden sondern fast völlig gestreckt. Die einzelnen Stränge eines *Faltblatts* können sowohl parallel als auch antiparallel angelagert sein. Die Stabilisierung dieser Konformation erfolgt ebenfalls über Wasserstoffbrücken, jedoch sind die an einer Brücke beteiligten Aminosäurereste nicht wie in Helices vier Positionen sondern beliebig weit in der Peptidkette voneinander entfernt.
α–Helices und β–Stränge sind durch unstrukturierte Bereiche der Peptidkette verbunden. Da diese unstrukturierten Bereiche in der Regel gekrümmte Konformationen aufweisen, werden sie auch *Schleifen*– oder *Loop*–Regionen genannt.
In der Regel erfolgt die Zuordnung der Sekundärstrukturelemente aus der auf-

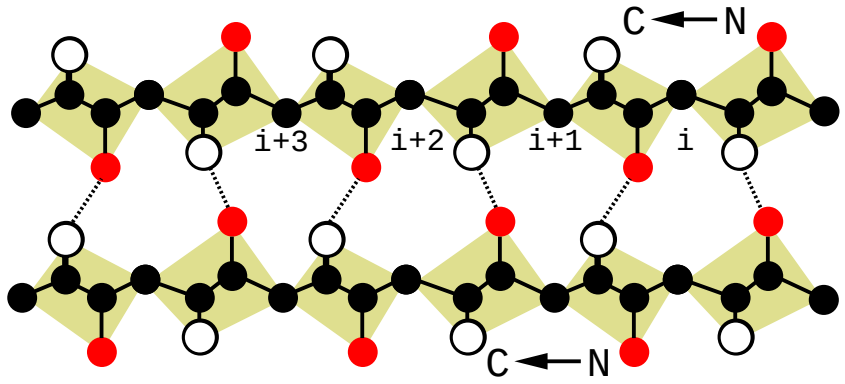


Abbildung 2.5: Paralleles β–Faltblatt mit Wasserstoffbrückenbindungen.

gelösten dreidimensionalen Struktur [173]. Ist die Proteinstruktur jedoch noch nicht aufgeklärt, so erlauben Zirkulardichroismus (*CD*) und Vibrationsspektroskopie die Messung des Anteils von Resten in Sekundärstrukturelementen. Falls genügend Protein aufgereinigt werden konnte, ist es unter Ausnutzung des nuklearmagnetische Resonanzeffektes (NMR) von bestimmten Atomtypen möglich, Distanzen zwischen bestimmten Aminosäureresten experimentell zu bestimmen [38, 368]. Aus den Distanzen können dann im nächsten Schritt die Sekundärstrukturelemente errechnet werden.

2.1.3 Tertiärstruktur

Die dreidimensionale Struktur eines Proteins wird Tertiärstruktur bezeichnet. Fast alle Proteine, zumindest alle Enzyme und regulatorischen Proteine, sind unter natürlichen Bedingungen zu globulären Strukturen dicht zusammengefaltet oder bestehen aus mehreren verbundenen Faltungseinheiten – den sogenannten *Domänen* –, die eben diese Eigenschaft aufweisen. Dabei liegen die hydrophoben Aminosäuren im Inneren und werden so von dem wäßrigen Lösungsmittel abgeschirmt. Die polaren Aminosäuren sind nach außen zum Lösungsmittel gerichtet und gehen Wechselwirkungen mit dem Wasser ein.
Die wesentlichen, die dreidimensionale Struktur stabilisierenden Interaktionen können wie folgt eingeteilt werden:

Elektrostatische Wechselwirkungen werden durch Ladungen oder Teilladungen von Aminosäureresten hervorgerufen:

- *Langreichweitige Wechselwirkungen* entstehen durch die elektrostatische Anziehung beziehungsweise Abstoßung elektrischer Ladungen oder Teilladungen.
- *Wasserstoffbrückenbindungen* werden zwischen dem sogenannten *Donor*, einem elektronegativen Atom mit kovalent gebundenem Wasserstoff, und dem sogenannten *Akzeptor*, einem weiteren elektronegativen

Atom, ausgebildet, wobei das Wasserstoffproton von dem Akzeptor angezogen wird.

- Als *Ionenbindung* wird die Anlagerung entgegengesetzt geladener Reste bezeichnet.

Van-der-Waals–Wechselwirkungen basieren auf Wechselwirkungen der Elektronenhüllen der beteiligten Atome. Zwischen diesen besteht eine attraktive Wechselwirkung, falls ihr Abstand größer als die Summe ihrer Van-der-Waals–Radien [271] ist, und eine repulsive Wechselwirkung, falls dieser Abstand unterschritten wird.

Disulfidbrücken sind kovalente Bindungen zwischen den Schwefelatomen von Cysteinen.

Hydrophobe Wechselwirkungen entstehen durch die Anlagerung apolarer Seitenketten. Durch diese Anlagerung werden Wassermoleküle von den apolaren Oberflächen verdrängt, so daß die freiwerdenden Wasserdipole untereinander Wechselwirkungen ausbilden können.

Scheinkräfte geben den entropischen Anteil wieder, der ein wesentlicher Faktor bei der Faltung ist.

Der jeweilige Energiebeitrag dieser Interaktionen für die Stabilität der Faltung variiert von Protein zu Protein, und er ist nur schwer zu bestimmen und noch schwerer allein anhand der Aminosäuresequenz vorherzusagen. Die Struktur eines Proteins ist nicht allein durch Wechselwirkungen zwischen den verschiedenen Aminosäuren bestimmt, sondern ein großer Teil der Struktur eines Proteins wird sicher bereits dadurch festgelegt, daß Proteine Kettenmoleküle sind. Insbesondere in Proteinstrukturvorhersageverfahren wird häufig der Tatsache, daß neben der Optimierung der verschiedenen Wechselwirkungen auch immer die Peptidkette noch erhalten bleiben muß, zu wenig Beachtung geschenkt [153].

2.1.4 Quartärstruktur

Die Quartärstruktur beschreibt die Struktur oligomerer Proteine, das heißt Proteine die aus zwei oder mehr Polypeptidketten aufgebaut sind. Als *Dimerisierung* (*Trimerisierung* etc.) wird die Anlagerung zweier (dreier etc.) Polypeptidketten bezeichnet. Die Dimerisierungsstellen der beteiligten Polypeptidketten sind – betrachtet man die einzelne Kette – in der Regel durch Bereiche mit hydrophoben Seitenketten gekennzeichnet. Die Hydrophobizität ist eine der treibenden Kräfte bei der Proteinfaltung. Die Hydrophobizität bewirkt, daß hydrophobe Seitenketten nicht an der Proteinoberfläche sondern im Proteininneren zu finden sind. Daraus ergibt sich, daß derartige hydrophobe Oberflächenbereiche nur aus der Dimerisierung heraus erklärbar sind und daher bei der Faltungsvorhersage Dimerisierungseffekte in jedem Falle miteinbezogen werden sollten.

2.2 Experimentelle Proteinstrukturbestimmung

Der folgende Abschnitt gibt einen kurzen Überblick über die experimentellen Methoden, die zur Aufklärung der dreidimensionalen Struktur von Proteinen eingesetzt werden. Die theoretische Proteinstrukturvorhersage, wie sie Thema dieser Dissertation ist, wird erst durch die Existenz und die breite Anwendung dieser Methoden möglich: Zum einen liefern die experimentell aufgeklärten Proteinstrukturen die Vorbilder beziehungsweise Modelle für die ähnlichkeitsbasierte Proteinstrukturvorhersage, die heute den wesentlichen Zugang zur Proteinstrukturvorhersage darstellt. Zum anderen ist die Menge der strukturaufgeklärten Proteine die Wissensbasis, die zur Ableitung von empirischen Parametern, als Trainingsmenge für wissensbasierte Systeme und als Testmenge zur Validierung der entwickelten Methoden dient.
Dieser Überblick soll außerdem verdeutlichen, daß theoretische Proteinstrukturvorhersagemethoden nicht nur da von Nutzen sind, wo experimentelle Methoden aufgrund ihres großen Aufwands nicht eingesetzt werden können (zum Beispiel bei der Analyse großer Datenbestände), sondern auch zur Lösung von Problemen beitragen, mit denen einige der experimentellen Methoden behaftet sind.
Gegenwärtig gibt es zwei etablierte experimentelle Verfahren, die dreidimensionale Struktur von Proteinen zu bestimmen, Röntgenkristallographie und Kernresonanzspektroskopie (*NMR*). Voraussetzung für beide Verfahren ist, daß das Protein in ausreichenden Mengen gereinigt zur Verfügung steht. Ein weiteres Verfahren, die Kryo–Elektronenmikroskopie, befindet sich gerade in der Entwicklung.

2.2.1 Röntgenkristallographie

Bei der Röntgenkristallographie wird die dreidimensionale Struktur des Proteins bestimmt, indem ein Proteinkristall mit Röntgenstrahlen bestrahlt und das resultierende Diffraktionsmuster gemessen wird. Das große Problem bei der Röntgenkristallographie besteht darin, wohlgeordnete Proteinkristalle ausreichender Größe zu züchten, so daß diese im Diffraktometer hochauflösende Beugungsmuster erzeugen. Die Züchtung der Kristalle ist schwierig, da Proteine große, kugel– oder ellipsoidförmige Objekte mit irregulären Oberflächen sind, die sich nur unter Einschluß zahlreicher Lösungmittelmoleküle zu einem dichten regulären Kristall packen lassen. Ist trotz dieser Schwierigkeiten die Züchtung eines Kristalls geglückt, so werden einige Röntgenstrahlen an den Elektronen der Atome gebrochen, und es ist möglich, ein Diffraktionsmuster aufzuzeichnen. Aus der Amplitude eines abgelenkten Röntgenstrahls (Intensität eines Punktes), der Wellenlänge der Röntgenquelle und der Phasenverschiebung des abgelenkten Röntgenstrahls können mittels Fouriertransformation die Koordinaten der Atome berechnet werden.
Bei dem Experiment geht jedoch die Information über die Phasenverschiebung verloren [40]. Zur Bestimmung der Phase haben Max Perutz und John Kendrew

die Methode der multiplen isomorphen Ersetzung entwickelt, bei der schwere Atome in den Kristall eingebaut werden, deren Position in dem Diffraktionsmuster eindeutig bestimmt werden können. Daraus läßt sich mit Hilfe der *Patterson*–Karte die Anordnung der schweren Atome und indirekt die Phase der an diesen Atomen abgelenkten Röntgenstrahlen bestimmen. Um daraus die Phase für die an Proteinatomen abgelenkten Röntgenstrahlen eindeutig bestimmen zu können, muß das gleiche Experiment mit einem zweiten Schwermetall–Proteinkomplex wiederholt werden [40]. Problematisch ist jeweils die Züchtung der Kristalle bzw. der Einbau der Schwermetalle in den Kristall, ohne diesen zu zerstören.
Das Phasenproblem ist jedoch wesentlich einfacher zu lösen, wenn bereits ein Modell für die zu lösende Struktur existiert. In diesem Falle kann das Phasenproblem mit der Methode der *molekularen Ersetzung* [343] rechnerisch gelöst werden. Dieses Modell kann sowohl aus NMR–Daten als auch aus dem Wissen über homologe Strukturen abgeleitet sein, die bereits strukturaufgeklärt sind. Die Suche nach diesen homologen Strukturen – insbesondere für die Fälle, wo diese Homologie nicht rein auf der Ebene der Primärstruktur erkannt werden kann – ist ebenso Inhalt dieser Dissertation wie die Verbesserung der aus den gefundenen homologen Strukturen abzuleitenden Modelle.

2.2.2 Kernresonanzspektroskopie

Das zweite experimentelle Verfahren zur Strukturaufklärung von Proteinen, die Kernresonanzspektroskopie (NMR), hat gegenüber der Röntgenkristallographie folgende Vorteile:

- Die Messung erfolgt in der Umgebung, in der sich Proteine auch in Organismen befinden, nämlich in wäßriger Lösung.
- Der aufwendige und manchmal auch erfolglose Arbeitsschritt der Kristallzüchtung entfällt.
- Es gibt in den Meßergebnissen keine Artefakte, die durch die Kristallpackung entstehen können.

Bei der NMR–Methode wird ausgenutzt, daß bestimmte Atomkerne (zum Beispiel ^{1}H, ^{13}C, ^{15}N und ^{31}P) ein magnetisches Moment oder *Spin* haben. Dieser Spin richtet sich in einem starken magnetischen Feld, wie es durch die Supraleiter ermöglicht wird, entlang dieses Feldes aus. Mittels einer Radiofrequenz können die Atomkerne jedoch kurzzeitig in einen angeregten Zustand gebracht werden. Beim Zurückfallen in den Gleichgewichtszustand emittieren die Atomkerne eine Radiofrequenz, die abhängig vom Atomtyp und der molekularen Umgebung dieses Atoms ist. Durch die Erweiterung der NMR–Methode auf zwei Dimensionen ist es möglich, sowohl Beziehungen zwischen über eine Folge von zwei oder drei kovalenten Bindungen verbundenen Wasserstoffatomen zu bestimmen (Korrelationsspektroskopie, *COSY*) als auch unter Ausnutzung des *nuklearen Overhauser*

Effektes (*NOE*) räumliche Nachbarschaften zwischen in der Aminosäuresequenz weit voneinander entfernten aber im Raum benachbarten (Abstand < 5Å) Aminosäuren zu messen. Aus den gemessenen Spektren lassen sich mit Hilfe der von Kurt Wüthrich entwickelten *sequentiellen Zuordnungstechnik* [368] Distanzbedingungen zwischen der Aminosäuresequenz zugeordneten Wasserstoffatomen beziehungsweise den dazugehörigen Resten ableiten. Aus diesen Distanzen können leicht die Sekundärstrukturelemente lokalisiert werden, da diese, wie oben beschrieben, durch typische geometrische Anordnungen charakterisiert sind. Die mit der NMR–Methode bestimmten dreidimensionalen Strukturen sind in der Regel nicht an allen Positionen eindeutig, da die verfügbaren Distanzbedingungen mehrere voneinander abweichende Konformationen zulassen. Dies ist jedoch nicht das den Einsatz der Methode limitierende Moment, das vielmehr darin besteht, daß die NMR–Methode zum gegenwärtigen Zeitpunkt auf Proteine mit höchstens etwas über zweihundert Aminosäuren begrenzt ist.
Ein Vergleich von mit kristallographischen und NMR–Methoden aufgeklärten Proteinstrukturen zeigt in der Regel Abweichungen, die im Rahmen der Meßgenauigkeit liegen [40] oder in der konformationellen Variabilität bestimmter Proteine begründet sind. In der Regel liegen die Unterschiede zwischen der NMR–Struktur und Kristallstruktur für die Rückgratatome im Proteininneren zwischen 0.16 und 0.79Å *RMS*–Abweichung (siehe Definition 3.4), in Einzelfällen kann diese Zahl allerdings bis zu 1.5Å betragen [134].

2.2.3 Kryo–Elektronenmikroskopie

Die Kryo–Elektronenmikroskopie [76] befindet sich noch im Entwicklungsstadium und ist in ihrer Handhabung noch sehr aufwendig. Daher wird sie nur dort eingesetzt, wo die beiden anderen Verfahren nicht angewandt werden können. Erste Erfolge sind zum Beispiel bei der Strukturaufklärung von Membranproteinen zu verzeichnen [213], die bislang nicht kristallisiert werden konnten. Die Struktur des Membranproteins Bakteriorhodopsin wurde zum Beispiel gerade in letztem Jahr mit einer Auflösung von 2.5Å kristallographisch aufgeklärt [275], wobei eine neue Technik zur Züchtung von Kristallen in Gittern von membranartigen Materialien angewendet wurde. Bei der Elektronenmikroskopie reicht es, Kristalle anzufertigen, die in einer Dimension nur eine oder wenige Molekülschichten dick sind. Ergebnis einer elektronenmikroskopischen Messung ist nicht nur ein Beugungsmuster sondern auch eine direkte Abbildung des Objektes, aus der die für die Interpretation des Beugungsmusters benötigte Phaseninformation abgeleitet werden kann. Die direkte Abbildung des Objektes ist möglich, da für Elektronen mittels Magnetfeldern die Funktion von optischen Linsen nachgebildet werden kann. Problematisch bei der Kryo–Elektronenmikroskopie sind neben der Tatsache, daß die Elektronen während der Messung die Probe zerstören, die Umgebungsbedingungen, unter denen die Messungen stattfinden. Die Messungen erfolgen im Hochvakuum an schockgefrorenen Proteinen, also unter Randbedingun-

gen, die eine spezielle Präparation der Proteine voraussetzen. Gegenwärtig liegt die Auflösung der Kryo–Elektronenmikroskopie in typischen Anwendungen [57] bei etwa 9Å und damit weit unter der Genauigkeit der Röntgenkristallographie. Da die Aufklärung der dreidimensionalen Struktur von Proteinen derart aufwendig ist, werden die Koordinaten insbesondere von Firmen aber auch von vielen Forschergruppen nicht oder nur mit großer Zeitverzögerung der Allgemeinheit zur Verfügung gestellt. Um die Aufklärung der Struktur zu dokumentieren, werden daher häufig nur stereographische Bilder der Proteinstruktur veröffentlicht. Dies hat dazu geführt, daß Verfahren entwickelt wurden, um aus diesen Bildern die Koordinaten berechnen zu können [161].

2.3 Proteinfaltung

Die meisten globulären Proteine falten in nativen Umgebungen spontan, das heißt ohne daß dazu zusätzliche äußere Einflüsse notwendig sind. Dies kann durch Experimente nachgewiesen werden, bei denen ein Protein durch Hitze oder `ph`–Veränderung denaturiert, das heißt entfaltet, wird. Nachdem die nativen Randbedingungen wiederhergestellt sind, kann man beobachten, daß die Proteine wieder ihre ursprüngliche Faltung annehmen. Auf der Basis dieser Experimente stellte Anfinsen [12, 13] bereits in den sechziger Jahren die Hypothese auf, daß die nativen Strukturen von Proteinen thermodynamisch stabile Zustände darstellen und sich damit gefaltete Proteine im Zustand der global minimalen *freien Energie* befinden.
Die Faltung eines Proteins erfolgt zeitlich im Bereich zwischen einigen Millisekunden und einigen Sekunden. Die Anzahl der möglichen Konformationen einer Proteinkette ist jedoch so groß, daß die Faltung zur nativen Struktur keinesfalls durch zufällige Suche im Konformationsraum ablaufen kann, wenn diese Konformation in dem beobachtenden Zeitraum gefunden werden soll. Die Diskrepanz zwischen der Faltungsgeschwindigkeit und der Anzahl der möglichen Konformationen wurde bereits 1969 von Cyrus Levinthal thematisiert [203] und ist aus diesem Grunde als *Levinthal Paradox* bekannt. Levinthal schloß aus seiner Beobachtung, daß die Proteinfaltung unter nativen Bedingungen entlang bestimmter Faltungswege (*folding pathways*) ablaufen muß.
Der Ablauf der Faltung sowohl *in vivo* als auch *in vitro* ist Gegenstand der aktuellen Forschung. Ruddon *et al.* diskutieren in einem Übersichtsartikel [294] die Proteinfaltung im endoplasmatischen Retikulum unter besonderer Berücksichtigung des Einflusses von *Chaperonen* und *Faltungskatalysatoren*. Chaperone sind Proteine, von denen angenommen wird, daß sie anderen Proteinen bei der Ausbildung der nativen Faltung helfen, indem sie Fehlfaltungen verhindern [138, 279]. Sie tun dies, indem sie ein zu frühes Einsetzen der Faltung [140] oder Interaktionen mit den vielen anderen Molekülen in der Umgebung des Ribosoms verhindern [92, 220], die zur Ausbildung falscher, nicht–funktioneller Strukturen führen. Anscheinend spielen dabei Chaperone in Eukaryonten für die Faltung eine größere

Rolle als in Prokaryonten [252]. Faltungskatalysatoren sind Proteine, zum Beispiel Isomerasen, die spezifische Faltungsschritte beschleunigen [138].
Die meisten Arbeiten zur Proteinfaltung sind daher im wesentlichen theoretischer Natur und versuchen die Proteinfaltung *in silico* unter Verwendung stark vereinfachter Modelle (Gitterabstraktionen, primitive Kraftfelder) zu simulieren [81]. Die dabei entwickelten Hypothesen über Faltungswege und frühe Faltungseinheiten (*early folding units*) [230] können dabei teilweise durch folgende Experimente belegt werden:

- Fluoreszenzspektroskopie ermöglicht die Beobachtung der Lösungsmittelzugänglichkeit von Tyrosin– und Tryptophanresten.
- Zirkulardichroismus (CD) und Vibrationsspektroskopie erlauben die Messung des durchschnittlichen Anteils bereits ausgebildeter Sekundärstrukturen.
- Mit der Methode der gepulsten Wasserstoffaustauschmarkierung kann die Zugänglichkeit der Stickstoffprotonen des Rückgrates protokolliert werden.

Der genaue Ablauf des Faltungsprozesses kann mit den heutigen experimentellen Methoden zur Strukturbestimmung nicht beobachtet werden, da diese nur die Beobachtung von Prozessen auf atomarer Ebene erlauben, die im Bereich von einigen Minuten bis Stunden ablaufen. Viele Wissenschaftler sind sogar der Meinung, daß dies im Detail nie oder zumindest nicht in absehbarer Zeit möglich sein wird [316], da dies die Entwicklung vollkommen neuer experimenteller Techniken erfordern würde.

Kapitel 3

Das Proteinstrukturvorhersageproblem

3.1 Motivation

Wie in Abschnitt 2.2 beschrieben ist die experimentelle Aufklärung der dreidimensionalen Struktur eines Proteins aufwendig und mit vielen Schwierigkeiten verbunden. Die Bestimmung der Sequenz ist dagegen experimentell – insbesondere auf DNA–Ebene – vergleichsweise einfach und weitgehend automatisiert. Diese Unterschiede spiegeln sich eindrucksvoll in den frei verfügbaren Datenbanken wieder (siehe Abbildung 3.1). In der Datenbank der aufgeklärten Proteinstrukturen

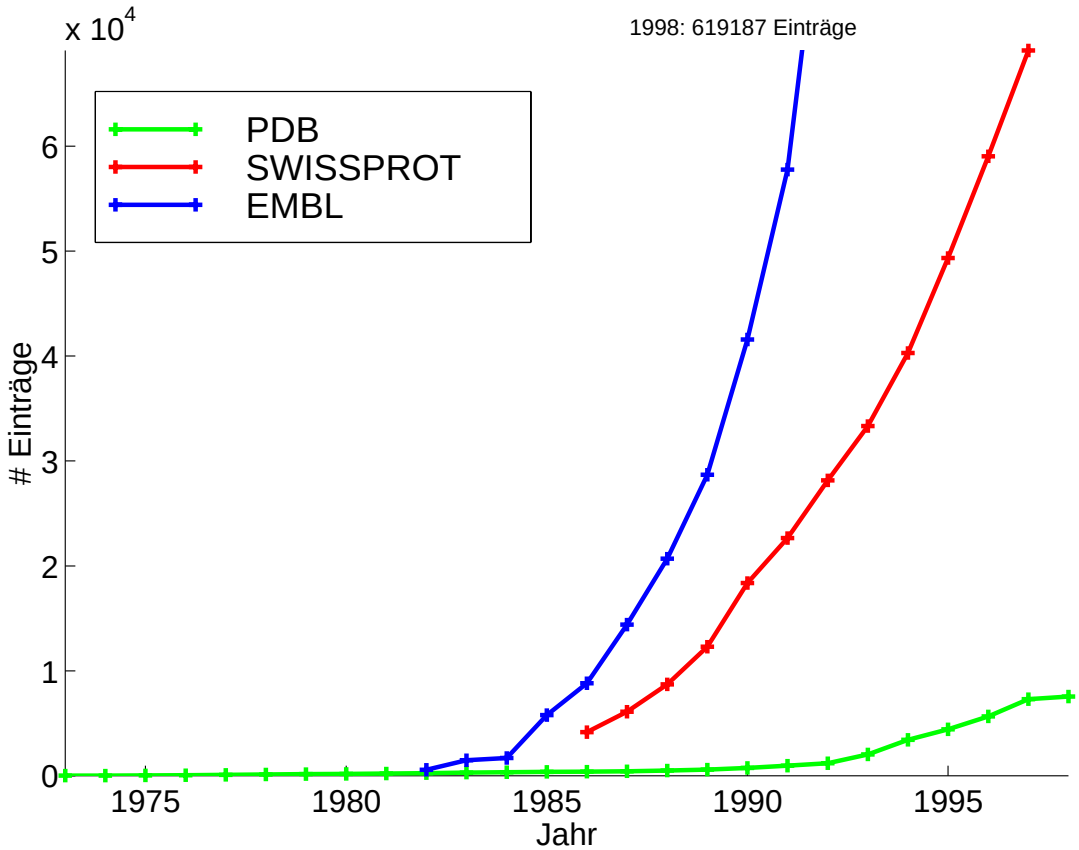


Abbildung 3.1: Wachstum der Proteinstrukturdatenbank (`PDB`), Proteinsequenzdatenbank (`SwissProt`) und Nukleotidsequenzdatenbank (`EMBL`).

(`PDB`) [29] gibt es derzeit 6.278 Einträge (1998). Im Gegensatz dazu enthält die öffentlich zugängliche Datenbank von Proteinsequenzen (`SwissProt`) [19] derzei-

tig 69.113 Sequenzen (Release 35, 12/97) und wächst zwischen den normalerweise halbjährlich veröffentlichten Versionen um durchschnittlich 15%. Die Nukleotidsequenzendatenbank des EMBL enthält gegenwärtig bereits 1.917.868 Nukleotidsequenzen (Release 53, 12/97). Es gibt keine verläßlichen Zahlen, wie viele Sequenzen und Strukturen die pharmazeutische Industrie und Biotechnologiefirmen in ihren nicht zugänglichen Datenbanken haben.
Auch die Schätzungen der Anzahl der Gene des menschlichen Genoms – mit dessen vollständiger Sequenzierung bis zum Jahr 2005 gerechnet wird – variieren zwischen 60.000 und 150.000 Genen [64]. Da die meisten dieser Gene Proteine codieren und neben der Sequenzierung des menschlichen Genoms auch die Genome weiterer Organismen sequenziert werden, ist mit einem weiteren rasanten Wachstum der Menge bekannter Proteinsequenzen zu rechnen, mit dem die experimentelle Strukturaufklärung (siehe 2.1.3) keinesfalls Schritt halten kann. Die Kenntnis der dreidimensionalen Struktur eines Proteins ist jedoch der grundlegende Schritt zum Verständnis seiner Funktion auf molekularem Niveau und damit auch die notwendige Voraussetzung zum zielgerichteten Wirkstoffentwurf [35, 36]. Daher wird das Proteinstrukturvorhersageproblem häufig auch als ein *Grand Challenge* Problem der Molekularbiologie bezeichnet. Gegenwärtig werden weltweit große Anstrengungen unternommen, diese Lücke zwischen bekannten Sequenzen und aufgeklärten Strukturen durch die Entwicklung von theoretischen, computerunterstützten Strukturvorhersagemethoden zu schließen beziehungsweise verringern. Theoretisch abgeleitete Strukturmodelle können dabei helfen, auch wenn sie nur Auskunft über Teilaspekte der Proteinstruktur geben, wie zum Beispiel die Faltungsklasse [228, 243, 260] oder ein Modell für das aktive Zentrum des Proteins [152, 381].
Theoretisch abgeleitete Strukturmodelle tragen jedoch nicht nur zum funktionellen Verständnis von nicht strukturaufgeklärten Proteinen bei sondern vereinfachen auch die experimentelle Strukturbestimmung. Bei Vorliegen eines theoretischen Strukturmodells wird durch die Methode der *molekularen Ersetzung (molecular replacement)* [343] bei der röntgenkristallographischen Strukturaufklärung der anderweitig zur Phasenbestimmung notwendige Arbeitsschritt der *multiplen isomorphen Ersetzung* [40] überflüssig.
Da sich dadurch die experimentelle Strukturaufklärung wesentlich vereinfacht und verkürzt, wird diese Methode heute bei allen Proteinen eingesetzt, bei denen eine Verwandtschaft zu bereits bekannten Proteinstrukturen bekannt ist oder vermutet wird. Wie Verwandtschaften zur Ableitung von Strukturmodellen verwendet werden können, beschreibt das Kapitel 3.3.

3.2 Formale Definition des Proteinstrukturvorhersageproblems

Die Problemstellung, aus der Sequenz der Aminosäuren eines Proteins seine dreidimensionale Struktur vorherzusagen, wird als das *Proteinstrukturvorhersageproblem* oder auch als die zweite Hälfte des genetischen Codes [185] bezeichnet. Ex-

perimente, bei denen die Zurückfaltung denaturierter Ribonucleasen in die native Konformation untersucht wurde, veranlaßten Anfinsen zur Aufstellung der bis heute weder bewiesenen noch widerlegten *thermodynamischen Hypothese*, daß sich Proteine in der nativen Konformation im Zustand der global minimalen *freien Energie* ΔG befinden [12, 13]. Als native Konformation eines Proteins wird dabei die Konformation verstanden, die das Protein unter Temperatur– und `ph`–Bedingungen annimmt, wie sie dort herrschen, wo das Protein normalerweise in einem Organismus gefunden wird.
Setzt man die Richtigkeit der thermodynamischen Hypothese voraus, so kann das Proteinstrukturvorhersageproblem formal wie folgt definiert werden:

Definition 3.1 (Proteinstrukturvorhersageproblem)

Gegeben:
- *eine Sequenz* $A =< a_i >$ *mit* $a_i \in \Sigma$, *wobei* Σ *das Alphabet der Einbuchstabencodes der* 20 *proteinogenen Aminosäuren ist.*
- *eine Menge von zulässigen Konformationen* $\mathcal{K}_A = \{K : A \longrightarrow I\!R^3\}$
- *die freie Energie* ΔG

Gesucht: *Eine Konformation* $K \in \mathcal{K}_A$ *für die gilt:*

$$\Delta G(K) = \min_{J \in \mathcal{K}_A} \Delta G(J)$$

So einfach diese Problemdefinition auf den ersten Blick erscheint, so schwierig, wenn nicht sogar unmöglich ist es, dieses Problem zu lösen.
Dies beginnt mit der Größe des Konfigurationsraums. Die Anzahl der möglichen dreidimensionalen Konformationen einer Proteinkette wächst exponentiell in der Anzahl der betrachteten Objekte (zum Beispiel einzelne Atome oder Aminosäuren). Diese wiederum ist abhängig von dem den Betrachtungen zugrunde liegenden Modell. Je genauer ein Modell die Wirklichkeit beschreiben soll, desto mehr Detailinformation muß es enthalten, desto komplexer und aufwendiger wird es. Die betrachteten Modelle reichen von reinen C_α– oder C_β–Modellen, über Modelle, die alle Proteinatome betrachten, bis hin zu Modellen, die sogar die das Protein umgebende Wasserhülle miteinbeziehen. Eine zulässige Konformation eines Proteins kann durch folgende Eigenschaften charakterisiert werden:

- Die Abstände und Winkel zwischen über eine oder mehrere chemische Bindungen verbundenen Atomen entsprechen den aus den chemischen Bindungslängen und –winkeln abgeleiteten Bedingungen.
- Es gibt keine sterischen Überlappungen von Atomen.
- Die Kette enthält keine topologischen Knoten. (Ausnahmen von dieser Regel sind nur wenige bekannt und sind abhängig von der Definition eines topologischen Knotens [214, 215].)

Das größere Problem stellt die zu minimierende Energiefunktion dar. Nach dem *zweiten thermodynamischen Hauptsatz* errechnet sich die freie Energie ΔG aus der Enthalpie ΔH, der absoluten Temperatur T und der Änderung der Entropie ΔS des zu beschreibenden Systems:

$$\Delta G = \Delta H - T\ \Delta S$$

ΔG wird auch als die *Gibbssche freie Energie* beziehungsweise *freie Enthalpie* bezeichnet. Die Entropie ist eine thermodynamische Zustandsgröße, die das Maß des mikrophysikalischen Unordnungszustands beschreibt, und errechnet sich aus der Anzahl gleichwertiger Anordnungen eines bestimmten Zustandes. Die Entropie eines Systems erreicht ihr Maximum im wahrscheinlichsten Zustand des Systems, der gleichzeitig der Zustand der maximalen Unordnung ist. Bezogen auf ein Proteinmolekül in wäßriger Lösung bedeutet das, daß bei der Ausbildung der nativen Konformation des Proteins die Unordnung der assoziierten Wassermoleküle zunimmt. Die Berechnung der Entropie einer Proteinstruktur ist nicht möglich, da dies die Kenntnis über die Verteilung aller gleichwertigen Anordnungen eines Zustandes des Proteins und des sie umgebenden Lösungsmittels erfordern würde. Näherungsweise wird versucht, die Entropie mit aufwendigen statistischen Methoden zu berechnen, die auf Ensembles von Strukturen rechnen.
Die Berechnung des enthalpischen Anteils der freien Energie der nativen Konformation kann approximativ durch Summation der verschiedenen Bindungs– und Wechselwirkungsenergien erfolgen. Diese umfassen unter anderem die Energiebeiträge von Wasserstoffbrücken, hydrophoben und elektrostatischen Wechselwirkungen aber auch Torsionswinkelenergien. Leider erfolgt die Stabilisierung einer Proteinstruktur nicht allein und manchmal nicht einmal überwiegend durch den enthalpischen Anteil der freien Energie.
Die Definition der freien Energie nach den Gesetzen der statistischen Mechanik (*Helmholtz freie Energie*) erfolgt durch eine sogenannte Zustandssumme, die die Summe der *Boltzmann*–Gewichte aller Energiestufen eines Systems ist [30]. Diese Zustandssumme kann jedoch nur für äußerst einfache Modellsysteme analytisch ausgedrückt und berechnet werden [188]. Für viele Systeme kann man näherungsweise die quantenmechanische Zustandssumme verwenden, die eine kontinuierliche Funktion ist. Zur Auswertung dieser Funktion hat jedoch die Integration über alle Freiheitsgrade des betrachteten Systems zu erfolgen, d.h. über $3N$ Freiheitsgrade, wobei N die Anzahl der Atome des Systems ist. Damit scheidet dieser Berechnungsweg für die in dieser Arbeit betrachteten Systeme (Proteine) natürlich aus. Realistisch gesehen kann er nur verwendet werden, um die Unterschiede in der freien Energie zwischen verwandten Systemen zu bestimmen, die über einfachste Änderungen ineinander überführt werden können [188].
In einer Übersicht [348] schätzt van Gunsteren 1990 die für die Simulation der Proteinfaltung in einem adäquaten Modell auf einem Hochleistungsrechner benötigte CPU–Zeit mit etwa $10^9 h$ ab und folgert, daß es bei kontinuierlich fortschrei-

tender Weiterentwicklung der Rechner in etwa hundert Jahren möglich sein wird, die in der Natur ablaufenden Prozesse zu simulieren.
Die *Helmholtz freie Energie* und die *Gibbssche freie Energie* beschreiben im wesentlichen das gleiche, nur daß die Gibbsche freie Energie als Summe von Enthalpie und Entropie definiert ist, während die Helmholtz freie Energie über die Boltzmann–gewichtete Zustandssumme über alle Energiezustände eines Systems definiert wird. Daher wird im folgenden immer vereinfachend von der freien Energie gesprochen.
Der Unterschied in der freien Energie zwischen dem gefalteten und ungefaltetem Zustand eines Proteins wird allgemein mit zwischen 40 $kJmol^{-1}$ [353] und 70 $kJmol^{-1}$ [70] angegeben. Dagegen schwankt die Gibbssche freie Energie sowohl des gefalteten als auch des ungefalteten Proteins über einen Temperaturbereich von 100 $°C$ um fast 3000 $kJmol^{-1}$ [70]. Zum Vergleich dazu tragen

- eine Wasserstoffbrückenbindung mit ungefähr -20 $kJmol^{-1}$ [353],
- eine Van-der-Waals–Wechselwirkung zwischen Atomen bei Anlagerung im bestmöglichen Abstand zwischen -2 und -4 $kJmol^{-1}$ und
- eine hydrophobe Wechselwirkung in Abhängigkeit von der angelagerten Oberfläche mit -0.1 $kJ\text{Å}^{-2}mol^{-1}$ [91]

zu der Stabilisierung der Struktur bei. Das bedeutet, daß der energetische Abstand zwischen der nativen Faltung und dem denaturierten Zustand eines Proteins kleiner ist als die Energiedifferenz, die für das Protein im nativen Zustand bei unterschiedlichen Temperaturen gemessen wird.
Für Computersimulationen und die dabei verwendeten Verfahren zur Berechnung der freien Energie eines Zustandes bedeutet dies, daß schon sehr kleine Fehler bei der Berechnung sich zu Fehlern aufsummieren können, die in der Größenordnung der freien Energie–Differenz zwischen dem gefalteten und ungefalteten Zustand liegt. Desweiteren resultiert aus den obigen experimentellen Meßergebnissen, daß neben der nativen Konformation eine Vielzahl weiterer nicht–nativer Faltungszustände existieren, die sich energetisch nur unwesentlich, von ihrer dreidimensionalen Gestalt jedoch grundlegend von der nativen Faltung unterscheiden. Dies macht das Proteinfaltungsproblem, wie es oben formuliert ist, zu einem Optimierungsproblem bei dem Millionen lokaler Minima existieren, die sich zusätzlich in ihrem Zielfunktionswert nur unwesentlich voneinander unterscheiden. Und dies gilt selbst dann, wenn die Energiefunktion genauestens berechnet werden könnte. Diese Problemeigenschaft und die Größe des zu betrachtenden Konformationsraums, läßt es äußerst unwahrscheinlich erscheinen, daß eine Suchprozedur, die sich notwendigerweise auf einen Ausschnitt des Konformationsraums beschränken muß, die Konformation mit global minimaler Energie finden kann.
Doch selbst unter der Annahme, daß es möglich ist, alle oben aufgeführten Probleme aus den Weg zu räumen, bleibt immer noch die Frage bestehen, ob die

native Konformation in allen Fällen auch wirklich die Konformation der minimalen freien Energie unter den durch die Temperatur und den ph–Wert vorgegeben Umgebungsbedingungen ist. Denn wie bereits erwähnt, wird heute allgemein davon ausgegangen, daß die Faltung eines Proteins schon bei seiner Synthese mit der Ausbildung sogenannter früher Faltungseinheiten beginnt. Dies kann bedeuten, daß frühzeitig lokale Faltungstrukturen eingenommen werden, deren Auflösung zur Erreichung des globalen Energieminimums notwendig sein kann, aber aufgrund der dabei zu überwindenden Energiebarriere vielleicht nicht möglich ist.
Zusammenfassend kann aus der obigen Diskussion gefolgert werden, daß das Proteinfaltungsproblem, wie es oben definiert ist, ein für Optimierungsstrategien schlecht konditioniertes Problem ist und eine Vorhersage der Proteinstruktur auf diesem Wege nicht möglich ist.

3.3 Homologie in Sequenz und Struktur

3.3.1 Sequenzen und Faltungsmotive

Die durchschnittliche Länge einer Domäne (siehe 2.1.3) eines Proteins in der Strukturdatenbank beträgt ungefähr 150 Aminosäurereste. Nimmt man die 20 proteinogenen Aminosäuren als Elementarbausteine, so lassen sich daraus 20^{150} verschiedene Aminosäureketten der Länge 150 zusammensetzen. Beschränkt man sich auf die Kombinationen, die sich untereinander durch eine Sequenzidentität (siehe Definition 3.3) kleiner als 20% auszeichnen, gibt es immer noch ungefähr 10^{38} nicht verwandte Möglichkeiten, eine Domäne der Länge 150 zu bilden [40]. In der Natur tritt jedoch nur ein äußerst geringer Anteil dieser theoretisch möglichen Kombinationen auf. Dies ist zum einen in der Entstehungsgeschichte der Proteine, der *Evolution*, begründet, zum anderen führen mit Sicherheit nicht alle Kombinationen zu einer stabilen dreidimensionalen Faltung. Doch selbst die konservative Annahme, daß nur jede millionste Möglichkeit eine 150 Aminosäurereste lange Domäne zu formen zu einer stabilen Faltung führt, läßt immer noch 10^{32} wirklich unterschiedliche Seitenkettenzusammenstellungen erwarten. Wie viele davon in der Natur wirklich vorkommen ist noch unbekannt.
Dieser enormen Größe des Sequenzraumes stehen auf der anderen Seite durch Beobachtungen belegte Schätzungen gegenüber, daß es in der Natur nur eine sehr begrenzte Anzahl von unterschiedlichen Proteinfaltungen gibt, wobei der Begriff, was eine Proteinfaltung ist und welche Strukturen damit zu einer Faltung zusammengefaßt werden, nicht klar definiert ist. Unter anderem aus diesem Grund reichen die Schätzungen über die Anzahl der in der Natur vorkommenden Proteinfaltungen von einigen hundert [358], weniger als 1.000 [58], ungefähr 7.000 [4], bis zu 8.000 [261] unterschiedlichen Proteinfaltungen. Die Schwankungen in diesen Abschätzungen sind sowohl durch die unterschiedlichen statistischen Ansätze als auch durch verschiedene Ähnlichkeitsmaße und unterschiedlichen Datenbestände zu erklären, die den entsprechenden Untersuchungen zugrunde gelegen haben.

Finkelstein und Ptitsyn versuchen in [99] anhand eines vereinfachten Modells zu ergründen, warum die Anzahl der Faltungsmuster globulärer Faltungen begrenzt ist.
Orengo *et al.* beobachten in ihren Analysen [260, 261] zudem, daß gemäß ihrer Klassifizierung von Faltungsmotiven einige Motive wesentlich häufiger auftreten als andere. Dies läßt sie vermuten, daß es sich bei den häufig auftretenden Faltungsmotiven um besonders stabile Faltungen handelt, die eventuell von einem gemeinsamen Vorfahren abstammen, obwohl diese gemeinsame Herkunft auf der Sequenzebene nicht mehr erkennbar ist [261].
Unabhängig welcher der obigen Abschätzungen man Vertrauen schenkt, kommen auf eine Proteinfaltung ungefähr 10^{27} wirklich unterschiedliche Möglichkeiten – das heißt mit Sequenzidentität geringer als 20% –, diese Faltung unter Verwendung der 20 proteinogenen Aminosäuren zu realisieren.

3.3.2 Sequenzhomologie und strukturelle Ähnlichkeit

Der Zusammenhang zwischen der Ähnlichkeit zweier Proteinsequenzen und der daraus resultierenden Ähnlichkeit der zugehörigen Strukturen wurde schon frühzeitig erkannt, nachdem die experimentelle Strukturaufklärung möglich geworden war [40]. Diese Erkenntnis wird heute in der *ähnlichkeitsbasierten Proteinstrukturvorhersage* ausgenutzt.

Definition 3.2 (Ähnlichkeitsbasierte Proteinstrukturvorhersage)
Bei der ähnlichkeitsbasierten Proteinstrukturvorhersage *wird für ein Protein, dessen dreidimensionale Struktur nicht experimentell aufgeklärt ist, ein Strukturmodell auf der Basis der Struktur eines strukturaufgeklärten Proteins erzeugt, zu dem zum Beispiel mit Vergleichsmethoden eine Verwandtschaft nachgewiesen werden kann.*

Üblicherweise erfolgt der Vergleich zweier oder mehrerer Proteinsequenzen in Form eines paarweisen beziehungsweise multiplen Alignments. In einem Sequenzalignment werden durch Einfügen von Leerstellen an beliebigen Stellen in den betrachteten Sequenzen gleiche oder ähnliche Aminosäuren der jeweiligen Sequenzen einander zugeordnet.

Definition 3.3 (Paarweises Sequenzalignment)
Seien $A =< a_1, \ldots, a_n >$ *und* $B =< b_1, \ldots, b_m >$ *zwei Folgen mit* $a_i, b_i \in \Sigma$, *wobei* Σ *das Alphabet der* 20 *proteinogenen Aminosäuren ist.*

- *Ein* Alignment *ist ein partieller injektiver Homomorphismus* $h : A \longrightarrow B$.
- *Die Menge* $\mathcal{H}(A, B) = \{h \mid h$ *partieller injektiver Homomorphismus* $h : A \longrightarrow B\}$ *bezeichne die Menge aller Alignments von* A *und* B.
- *Die Menge* $M_h(A, B) = \{a_i \in A \mid h(i) \neq \emptyset \wedge a_i = b_{h(i)}\}$ *ist die Menge der* Identitäten *oder* Matches.

- *Die Menge* $N_h(A,B) = \{a_i \in A \,|\, h(i) \neq \emptyset \wedge a_i \neq b_{h(i)}\}$ *ist die Menge der* Nichtübereinstimmungen *oder* Mismatches.
- *Die Menge* $D_h(A,B) = \{a_i \in A \,|\, h(i) = \emptyset\}$ *ist die Menge der* Deletionen.
- *Die Menge* $I_h(A,B) = \{b_j \in B \,|\, h^{-1}(j) = \emptyset\}$ *ist die Menge der* Insertionen.
- *Eine konsekutive Folge von Insertionen bzw. Deletionen wird im folgenden als* Gap *bezeichnet. Sei* $G_h(A,B)$ *die Menge der Gaps in dem Alignment* h.
- *Die* Sequenzidentität $Id_h(A,B)$ *zweier Sequenzen* A *und* B *bezüglich eines Alignments* h *ist definiert als*

$$Id_h(A,B) = \frac{|M_h(A,B)| * 100\%}{min\{|A|,|B|\}}$$

```
1tpp _000: ---------------IVGGYTCGANTVPYQVSL--NSGYHFCGGSLINSQWVVSAAHCYKSGIQVR----
1mctA_000: ---------------IVGGYTCAANSIPYQVSL--NSGSHFCGGSLINSQWVVSAAHCYKSRIQVR----
1trnA_000: ---------------IVGGYNCEENSVPYQVSL--NSGYHFCGGSLINEQWVVSAGHCYKSRIQVR----
1bit _000: FVLLIGAAFATEDDKIVGGYECKAYSQAHQVSL--NSGYHFCGGSLVNENWVVSAAHCYKSRVEVR----
1cgiE_000: CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTTSDVV----
2gmt _000: CGVPAIQPVLSGLUUIVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTTSDVV----
1hneE_000: ---------------IVGGRRARPHAWPFMVSLQLRGG-HFCGATLIAPNFVMSAAHC-VANVNVRAVRV
1sgt _000: ---------------VVGGTRAAQGEFPFMVRLSMG-----CGGALYAQDIVLTAAHCVSGSGNNTSITA
score_000:                |V|G | |||  |||V L       CG||L ||| V||A|HC  || |  || |

1tpp _070: L-GEDNINVVEGNEQFISASKSIVHPSYNSNTLNNDIMLIKLKSAASLNSRVASISLPTSCA--SAGTQC
1mctA_070: L-GEHNIDVLEGNEQFINAAKIITHPNFNGNTLDNDIMLIKLSSPATLNSRVATVSLPRSCA--AAGTEC
1trnA_070: L-GEHNIEVLEGNEQFINAAKIIRHPQYDRKTLNNDIMLIKLSSRAVINARVSTISLPTAPP--ATGTKC
1bit _070: L-GEHNIKVTEGSEQFISSSRVIRHPNYSSYNIDNDIMLIKLSKPATLNTYVQPVALPTSCA--PAGTMC
1cgiE_070: VAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTC
2gmt _070: VAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTC
1hneE_070: VLGAHNLSRREPTRQVFAVQRIFEDG-YDPVNLLNDIVILQLNGSATINANVQVAQLPAQGRRLGNGVQC
1sgt _070: TGGVVDLQ--SGAAVKVRSTKVLQAPGYNGTG--KDWALIKLAQP------INQPTLKIATTTAYNQGTF
score_070:   G  |||  |||   |||||||| | ||| |  |D  |||L| |      ||   L  | |   |  |

1tpp _140: LISGWGNTKSSGTSYPDVLKCLKAPILSDSSCKSAYPGQ-ITSNMFCAGYLE-GGKDSCQGDSGGPVVCS
1mctA_140: LISGWGNTKSSGSSYPSLLQCLKAPVLSNSSCKSSYPGQ-ITGNMICVGFLQ-GGKDSCQGDSGGPVVCN
1trnA_140: LISGWGNTASSGADYPDELQCLDAPVLSQAKCEASYPGK-ITSNMFCVGFLE-GGKDSCQGDSGGPVVCN
1bit _140: TVSGWGNTMSSTAD-SNKLQCLNIPILSYSDCNNSYPGM-ITNAMFCAGYLE-GGKDSCQGDSGGPVVCN
1cgiE_140: VTTGWGLTRYTNANTPDRLQQASLPLLSNTNCKKYWGTK-IKDAMICAG--A-SGVSSCMGDSGGPLVCK
2gmt _140: VTTGWGLTRYUUANTPDRLQQASLPLLSNTNCKKYWGTK-IKDAMICAG--A-SGVSSCMGDSGGPLVCK
1hneE_140: LAMGWGL-LGRNRGIASVLQELNVTVVT-SLCRR---------SNVCTLVRG-RQAGVCFGDSGSPLVCN
1sgt _140: TVAGWGANREGGSQQRYLLK-ANVPFVSDAACRSAYGNELVANEEICAGYPDTGGVDTCQGDSGGPMFRK
score_140:  ||GWG || ||||    L|  |||||||| C|| |||| |||| |C||| | || ||C|GDSG|P|  |

1tpp _210: GK-----LQGIVSWG-SGCAQKNKPGVYTKVCNYVSWIKQTIASN-  Trypsin        (Rind)
1mctA_210: GQ-----LQGIVSWG-YGCAQKNKPGVYTKVCNYVNWIQQTIAAN-  Trypsin     (Schwein)
1trnA_210: GQ-----LQGVVSWG-DGCAQKNKPGVYTKVYNYVKWIKNTIAANS  Trypsin      (Mensch)
1bit _210: GE-----LQGVVSWG-YGCAEPGNPGVYAKVCIFNDWLTSTMAS-Y  Trypsin       (Fisch)
1cgiE_210: KNG-AWTLVGIVSWG-SSTCSTSTPGVYARVTALVNWVQQTLAAN-  Chymotrypsin (Mensch)
2gmt _210: KNG-AWTLVGIVSWG-SSTCSTSTPGVYARVTALVNWVQQTLAAN-  Chymotrypsin   (Rind)
1hneE_210: G-----LIHGIASFVRGGCASGLYPDAFAPVAQFVNWIDSIIQ---  Elastase     (Mensch)
1sgt _210: DNADEWIQVGIVSWG-YGCARPGYPGVYTEVSTFASAIASA-ARTL  Trypsin        (Pilz)
score_210: ||| |||  G||S|| ||||||| P|||| V||||| | || | |   Konsensus
```

Abbildung 3.2: Multiples Alignment verschiedener Serinproteasen (Hydrolasen) aus verschiedenen Organismen [226].

Abbildung 3.2 zeigt ein multiples Alignment verschiedener Serinproteasen (Hydrolasen) aus verschiedenen Organismen. Die zu einem Protein gehörigen Zeilen

werden durch den jeweiligen PDB–Code identifiziert. Die Proteinnamen sind im letzten Block des multiplen Alignments angegeben.
Die Konsensuszeile hebt die konservierten Aminosäuren durch eine Wiederholung der konservierten Aminosäure hervor, wenn sie in allen Sequenzen konserviert ist, oder durch einen senkrechten Strich, wenn eine Aminosäure in der Mehrheit der Sequenzen an dieser Position zu finden ist. Bei den betrachteten Proteinen handelt es sich um Serinproteasen, die andere Proteine an klar definierten Positionen enzymatisch spalten. Der Vergleich der Sequenzen in Form des Alignments zeigt, daß die für die Funktion der Serinproteasen wesentlichen Aminosäuren – ein Histidin (im Alignment an Position 56), eine Asparaginsäure (105) und ein Serin (202) – in allen Serinproteasen konserviert sind. Diese drei Aminosäuren bilden die für die Serinprotease typische *katalytische Triade* und liegen in den jeweiligen dreidimensionalen Strukturen nah beieinander, obwohl sie in der Peptidkette weit voneinander entfernt liegen. Die Sustratspezifität entsteht durch die Unterschiede in den Aminosäureresten, die die Spezifitätstasche um die katalytische Triade bilden und nicht konserviert sind.
Ein mögliches Maß zum Vergleich zweier Strukturen stellt die Wurzel der mittleren Abweichungsquadrate (*root mean square deviation*, *RMS*) [281] der in einer optimalen Superposition der Strukturen einander zugeordneten Aminosäurereste dar. In der Regel wird die Abweichung auf Basis der Koordinaten der C_α–Koordinaten berechnet. Die Berechnung dieses Wertes beinhaltet implizit die Suche nach der größten gemeinsamen Punktmenge (*largest common point set problem*) der zwei zu vergleichenden Strukturen. Dieses Problem ist *NP*–vollständig und bleibt auch dann *NP*–vollständig, wenn man nicht Punktmengen sondern im dreidimensionalen Raum eingebettete Graphen betrachtet [2]. Daher wird der *RMS* in der Regel bezüglich eines vorgegebenen Alignments bestimmt, das die zu superpositionierenden Punktepaarungen vorgibt.

Definition 3.4 (Wurzel der mittleren Abweichungsquadrate(RMS))
Seien $A =< a_1, \ldots, a_n >$ *und* $B =< b_1, \ldots, b_m >$ *die Sequenzen zweier strukturaufgeklärte Proteinstrukturen und sei* h *ein Alignment von* A *mit* B. *Seien* $C_\alpha^X(i) = (x_i, y_i, z_i)$ *die kartesischen Koordinaten des* i*–ten* $C_\alpha(i)$ *Atoms der Struktur* X. *Dann ist die* Wurzel der mittleren Abweichungsquadrate *bezüglich des Alignments* h $RMS_h(A, B)$ *definiert als:*

$$RMS_h(A,B) = \min_{t \in \mathbb{R}^3, R \in \mathbb{R}^3 \times \mathbb{R}^3} \sqrt{\frac{\sum_{i:h(i)\neq\emptyset} \|C_\alpha^A(i) - (t + R * C_\alpha^B(h(i)))\|^2}{|\{i|h(i) \neq \emptyset\}|}},$$

wobei $\|.\|$ *die euklidische Abstandsnorm im* $\mathbb{R}^3$ *ist.*

Im folgenden ist mit *RMS* immer der RMS_h bezüglich eines Alignments h gemeint, das aus dem jeweiligem Zusammenhang eindeutig bestimmt ist. Ist die Zuordnung der zu superpositionierenden Koordinaten bekannt, kann die Berechnung der optimalen Superposition effizient mit Eigenwertmethoden [96, 171, 172] oder der konjugierten Gradientenmethode [224] erfolgen.

		1mctA	1trnA	1bit	1cgiE	2gmt	1hneE	1sgt
1tpp	Id[%]	81	75	65	44	44	33	33
	RMS[Å]	0.5	0.6	1.6	2.0	1.8	2.7	3.2
1mctA	Id[%]		78	65	45	45	35	33
	RMS[Å]		0.4	1.5	2.0	1.8	2.6	3.1
1trnA	Id[%]			63	44	44	34	29
	RMS[Å]			1.6	2.0	1.9	2.6	3.1
1bit	Id[%]				40	40	34	33
	RMS[Å]				2.4	1.9	2.8	3.3
1cgiE	Id[%]					98	32	32
	RMS[Å]					0.9	2.6	3.1
2gmt	Id[%]						32	32
	RMS[Å]						2.4	3.0
1hneE	Id[%]							25
	RMS[Å]							3.1

Tabelle 3.1: Sequenzidentität Id versus Strukturähnlichkeit$R$: Serinproteasen identifiziert durch ihre PDB–codes werden paarweise verglichen. Das Bezugsalignment sowohl für die Berechnung der Sequenzidentität Id als auch des RMS ist dem multiplen Alignment aus Abbildung 3.2 entnommen.

Der Vergleich in Tabelle 3.1 zeigt, daß der RMS mit steigender Sequenzidentität fällt und das Rückgrat der Strukturen (hier repräsentiert durch die C_α–Atome, die bei der Berechnung des RMS verwendet wurden) nahezu identisch ist, wenn die Sequenzidentität die 70%–Marke übersteigt. Aber auch wenn die Sequenzidentität sich der 25%–Marke nähert (vgl. *1hneE vs. 1sgt*), ist der RMS der optimalen Struktursuperposition klein (hier weniger als 3.1Å). Abbildung 3.3 zeigt die aus dem multiplen Alignment aus Abbildung 3.2 resultierende Superposition der C_α–Atome aller Serinproteasen auf die Struktur des Trypsins vom Rind. In Ergänzung zum RMS–Wert zeigt diese Abbildung, daß die strukturbestimmenden Elemente im Kernbereich der Proteine hoch konserviert sind. Das grobe Baugerüst (*Scaffold*) ist also identisch. Unterschiede bestehen im wesentlichen in den dem Lösungsmittel zugänglichen Schleifenbereichen. Diese Beobachtung deckt sich mit den Ergebnissen von Chothia und Lesk, die 1986 den Zusammenhang der Sequenzähnlichkeit und der strukturellen Ähnlichkeit der Kernbereiche der Proteine untersucht und quantifiziert haben [59].

Abbildung 3.4 zeigt den Zusammenhang zwischen der Sequenzhomologie und der strukturellen Ähnlichkeit in einer Analyse einer repräsentativen Menge und aller dazu homologen bekannten Proteinstrukturen. Eine erste umfassende Analyse dieses Zusammenhangs wurde 1991 von Sander und Schneider [310] veröffentlicht.

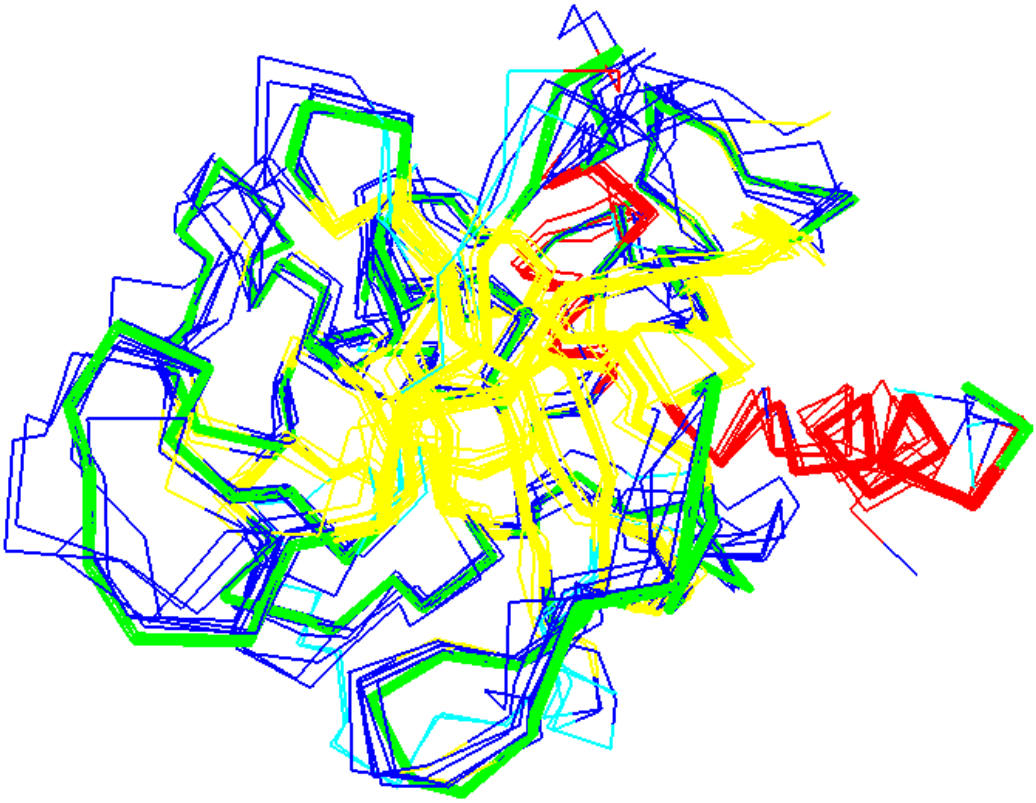


Abbildung 3.3: Strukturelle Superposition der Serinproteasen gemäß dem multiplen Alignment aus Abbildung 3.2. Bezugsstruktur ist das Trypsin vom Rind (*1tpp*, dicker Linienzug). Die Sekundärstrukturelemente sind farbig hervorgehoben (Helices in rot, Stränge in gelb).

Ein Ergebnis dieser Analyse war, daß die Sekundärstrukturelemente zu mehr als 70% identisch sind, wenn die Sequenzidentität zweier Proteine über 30% beträgt. Das bedeutet, daß in diesen Fällen von der Ähnlichkeit der Sequenzen auf die Ähnlichkeit der Strukturen geschlossen werden kann.

Die bei dieser Untersuchung entstandene Datenbank HSSP (*Database of homology derived structure*) [310] wird von den Autoren seitdem fortlaufend aktualisiert und enthält zu jeder bekannten Proteinstruktur ein multiples Alignment der homologen Proteinsequenzen zu der Sequenz des jeweiligen Proteins.

Die in Abbildung 3.4 gezeigten Werte basieren auf den multiplen Alignments der HSSP–Datenbank von 1996. Die Repräsentativmenge [148] ist so ausgewählt, daß die Sequenzidentität eines jeden Proteinpaares dieser Menge kleiner als 25% ist.

Bei der Analyse wurden nur die Proteine betrachtet, deren Struktur experimentell aufgeklärt ist. Desweiteren wurden alle Daten eliminiert, die offensichtlich auf Fehlern in den verwendeten Datenbanken (zum Beispiel auch Fehler im *chain*

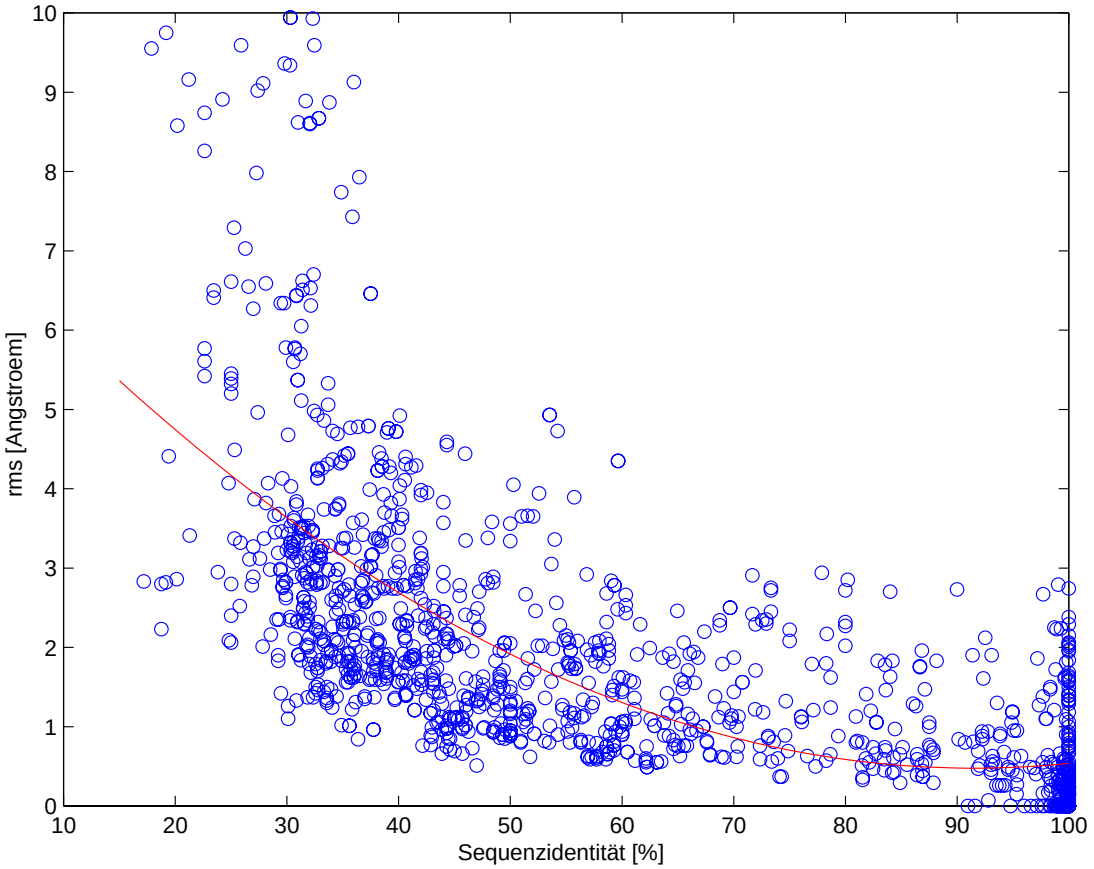


Abbildung 3.4: Analyse des Zusammenhangs von Sequenz– und Strukturähnlichkeit auf Basis einer repräsentativen Menge (`hobohm_96_25`, 488 Proteine) und der dazugehörigen homologen Proteine. Die Sequenzähnlichkeit ist gemessen in Prozent Sequenzidentität, die Strukturähnlichkeit ist gegeben als *RMS*[Å], beides bezüglich der in `HSSP` [310] gegebenen multiplen Alignments (1415 paarweise Alignments).

tracing) zurückzuführen sind. Da die Analyse auf die Proteinpaare eingeschränkt wurde, für die mindestens 50 Positionen aliniert sind, entfällt hier die in [310] diskutierte Längenabhängigkeit des Schrankenwertes zur Erkennung homologer Sequenzen. Im Unterschied zu Schneider und Sander dient hier der RMS_h der optimalen Superposition der Strukturpaare bezüglich des gegebenen Alignments h als Vergleichskriterium. Aus Abbildung 3.4 wird deutlich, daß selbst bei einer Sequenzidentität von 100% *RMS*–Abweichungen von fast 3Å möglich sind. In diesen Fällen handelt es sich in der Regel um verschiedene Strukturmodelle für ein und dasselbe Protein, die in verschiedenen Labors, mit verschiedenen Methoden, unter verschiedenen Randbedingungen oder mit unterschiedlicher Genauigkeit aufgeklärt worden sind.

In Ergänzung zu den Untersuchungen von Sander und Schneider zeigt Abbildung 3.4, daß die zu erwartende *RMS*–Abweichung mit sinkender Sequenzidentität zunimmt. Das bedeutet jedoch keinesfalls, daß die strukturelle Abweichung bei niedriger Sequenzidentität automatisch hoch ist. Im Gegenteil gibt es viele Strukturpaare, deren *RMS*–Abweichung um die 2Å liegt, obwohl sie nur eine Sequenzidentität unter 40% aufweisen. Mit abnehmender Sequenzidentität wird es nur wesentlich schwieriger, diese strukturelle Ähnlichkeit ohne Kenntnis der Strukturen zuverlässig vorherzusagen. Für den besonders schwierigen Bereich von Sequenzidentitäten geringer als 30% verwenden Sander und Schneider daher den Begriff der *twilight zone*, da hier eine strukturelle Ähnlichkeit zwar möglich ist, aber bei einem reinen Vergleich der Sequenzen nicht mehr zuverlässig erkannt werden kann.

Darüber hinaus zeigt Abbildung 3.4, daß im Bereich über 30% Sequenzidentität strukturelle Homologien zwar eindeutig erkannt werden können, aber Mißalignments trotzdem zu relativ hohen *RMS*–Abweichungen führen können. Ziel der in dieser Arbeit vorgestellten Methode ist es, auf der einen Seite die zuverlässige Vorhersage struktureller Ähnlichkeiten auf den Bereich unter 30% Sequenzidentität auszudehnen und zum anderen die Qualität der Alignments in dem Bereich darüber zu verbessern.

Nicht immer jedoch impliziert eine hohe Sequenzidentität auch eine strukturelle Ähnlichkeit. In einen Experiment ist es Reagan *et al.* [72] gelungen, durch zielgerichteten Austausch von weniger als 50% der Aminosäuren ein vorwiegend aus β–Strängen bestehendes natürliches Protein in ein künstliches Protein mit den typischen Eigenschaften eines nativen Proteins umzuwandeln, das die Faltung eines Vier–Helix–Bündels annimmt.

Kurze Sequenzfragmente (zum Beispiel einige Oktamere) können in nativen Proteinstrukturen sogar bei vollständiger Sequenzidentität in verschiedenen Sekundärstrukturkonformationen auftreten [331]. Dies zeigt, daß die Konformation eines Proteinfragmentes nicht allein durch lokale Eigenschaften festgelegt ist. Die auf Sequenzhomologie basierende Strukturvorhersage ist daher nur dann zuverlässig anwendbar, wenn sich die (Sequenz–)Ähnlichkeit über große Bereiche der untersuchten Sequenzen erstreckt und nicht nur auf lokale Regionen begrenzt ist.

Vergleichbare Beispiele finden sich auch in der Menge der nativen Proteinstrukturen. So haben die Kalzium bindenden Proteine untereinander eine Sequenzidentität von über 30% [272]. Da die Identitäten zudem über die gesamte Länge der Proteine verteilt sind, würde man nach den empirischen Regeln der vergleichenden Modellierung auch auf die Ausbildung ähnlicher Strukturen schließen. Eine Betrachtung der experimentell aufgeklärten Strukturen zeigt jedoch gravierende Unterschiede zwischen den Strukturen der Calcium bindenden Proteine, die sich durch eine *RMS*–Abweichung von ungefähr 10Å für die optimale strukturelle Superposition bestätigt [272].

Für die Proteinstrukturvorhersage bedeuten diese Ausnahmen, daß neben den benutzten Bewertungskriterien auch immer weitere biologische oder biochemische Informationen mit einbezogen werden sollten, um Vorhersagen zu untermauern.

3.3.3 Strukturelle Ähnlichkeit ohne signifikante Sequenzhomologie

Nach den empirischen Regeln der ähnlichkeitsbasierten Strukturvorhersage wird von einer nicht signifikanten Sequenzhomologie gesprochen, wenn der Prozentsatz der Identitäten für zwei Sequenzen in einem optimalen Alignment die Marke von 30% beziehungsweise 25% unterschreitet und man sich damit in der sogenannten *twilight zone* befindet. Wenn man zufällige Sequenzen vergleicht, die die für Proteine typische Aminosäurezusammensetzung haben, so ist ohne Optimierung durch Einfügen von Insertionen und Deletionen bereits eine Sequenzidentität von 6% zu erwarten [84]. Sequenzalignmentverfahren, die Insertionen und Deletionen erlauben und die Sequenzidentität als Bestandteil der Kostenfunktion verwenden, heben diese untere Grenze der für zufällige Sequenzen zu erwartenden Sequenzidentität beträchtlich an, so daß bereits bei Alignments mit einer Sequenzidentität unter 25% nicht mehr auf eine evolutionäre Verwandschaft der alinierten Sequenzen geschlossen werden kann [84].
Daß Proteine, deren Sequenzähnlichkeit in diesem Bereich liegt, dennoch signifikante strukturelle Ähnlichkeiten aufweisen können, ist spätestens durch den systematischen Vergleich von Proteinstrukturen mit der Hilfe von Computermethoden belegt. Diese Methoden vereinfachen nicht nur den visuellen Vergleich von Proteinstrukturen durch die optimale Superposition, sondern erlauben auch den systematischen paarweisen Vergleich der mittlerweile recht großen Menge von experimentell bestimmten Proteinstrukturen.
Methoden zum strukturbezogenen Vergleich von Proteinen sind seit geraumer Zeit Gegenstand der Forschung. Die Verfahren reichen von dem Vergleich der Sekundärstrukturzusammensetzung über das Finden maximaler Teilstrukturen gleicher Topologie [129, 183] bis hin zur Berechnung gemeinsamer Teilstrukturen, für die der *RMS* der optimalen Superposition einen vorgegebenen Grenzwert nicht überschreitet. In der Regel werden die Proteine dabei als starre Objekte betrachtet (*rigid body superposition*). Die Methoden der letzten Kategorie unterteilen sich wiederum in Verfahren, die an der Topologie der Proteinkette festhalten – wie zum Beispiel die Programme `SSAP` [263, 339], `Dali` [150] und `VAST` [115] – und Methoden wie `SARF` [7, 8], die auch die Aufdeckung strukturell ähnlicher Teilmotive erlauben, deren topologische Anordnung nicht erhalten ist [3, 257].
Gemeinsam ist allen diesen Verfahren, daß sie als Ergebnis eine Abbildung oder im die Topologie erhaltenden Fall ein Alignment zwischen den Aminosäureresten der verglichenen Strukturen liefern. Ist diese Abbildung bekannt, so kann – wie bereits im Abschnitt 3.3.2 beschrieben – die optimale Superposition der Strukturen berechnet und so ein visueller Vergleich der Proteine ermöglicht werden. Zur Bestimmung der einander zuzuordnenden Positionen kommen die unterschied-

lichsten Methoden zum Einsatz:

- Methoden aus dem Bereich der Bildverarbeitung, zum Beispiel *geometric hashing* [3, 257],
- Alignment struktureller Eigenschaften [67, 222] beziehungsweise von 3D–-Profilen [333] mittels dynamischer Programmierung oder anderer einfacher Vergleichsverfahren mit oder ohne anschließender Verfeinerung,
- Vergleich intramolekularer Abstände [150, 208],
- genetische Algorithmen [223],
- doppelte dynamische Programmierung [263, 339] und
- durch Suche nach gleichzeitig superpositionierbaren Sekundärstrukturelementen mit Clique–Algorithmen und mit anschließender Erweiterung auf Nichtsekundärstrukturpositionen [7, 8, 115].

Diese Methoden haben zur Aufdeckung entfernter evolutionärer Verwandtschaften geführt, die auf Sequenzebene nicht nachweisbar sind, aber Rückschlüsse auf unerwartete funktionelle Eigenschaften ermöglichen. Die entscheidende Weiterentwicklung der letzten Jahre besteht darin, daß die Methoden nicht mehr nur den strukturellen Vergleich einzelner Proteine, sondern die systematische Analyse aller bekannten Proteinstrukturen erlauben [6, 115]. Die Ergebnisse dieser Analysen sind für die Art der in dieser Arbeit entwickelten Proteinstrukturvorhersagemethoden von zentraler Bedeutung. Zum einen belegen sie, daß strukturelle Ähnlichkeiten auch jenseits hoher Sequenzidentität zu entdecken sind und zu Erkenntnissen über die Funktion eines Proteins beitragen können. Zum anderen können durch den strukturellen Vergleich die strukturbestimmenden Bestandteile durch Analyse der Gemeinsamkeiten verwandter Strukturen ermittelt werden.
Insbesondere zeigen die systematischen Strukturanalysen, daß es strukturelle Ähnlichkeiten zwischen Proteinen ohne signifikante Sequenzhomologie gibt, insbesondere auch für Proteine aus der sogenannten *twilight zone* mit einer Sequenzidentitätsrate geringer als 30% [6]. Wie bei einem Vergleich auf Sequenzebene gibt es auch beim Strukturvergleich signifikante und zufällige Ähnlichkeiten. Zur Abgrenzung signifikanter Ähnlichkeiten definieren Alexandrov und Fischer [7] einen Ähnlichkeitswert, den sogenannten *zscore* (siehe Definition 4.4), der die Ähnlichkeit zweier Proteine A und B in Bezug zur Ähnlichkeit von Protein A zu allen anderen Proteinen einer Repräsentativmenge setzt, die aus 320 nicht homologen beziehungsweise niederhomologen Proteinen besteht [101]. Die Ähnlichkeit wird dabei über die Anzahl der superpositionierbaren Positionen definiert, da die verwendete `SARF2`–Methode nur Strukturalignments liefert, die mit einem *RMS* kleiner als ein Grenzwert superpositionieren. Eine Strukturähnlichkeit wird von ihnen als signifikant angesehen, wenn der *zscore* größer als 3.0 ist, daß heißt

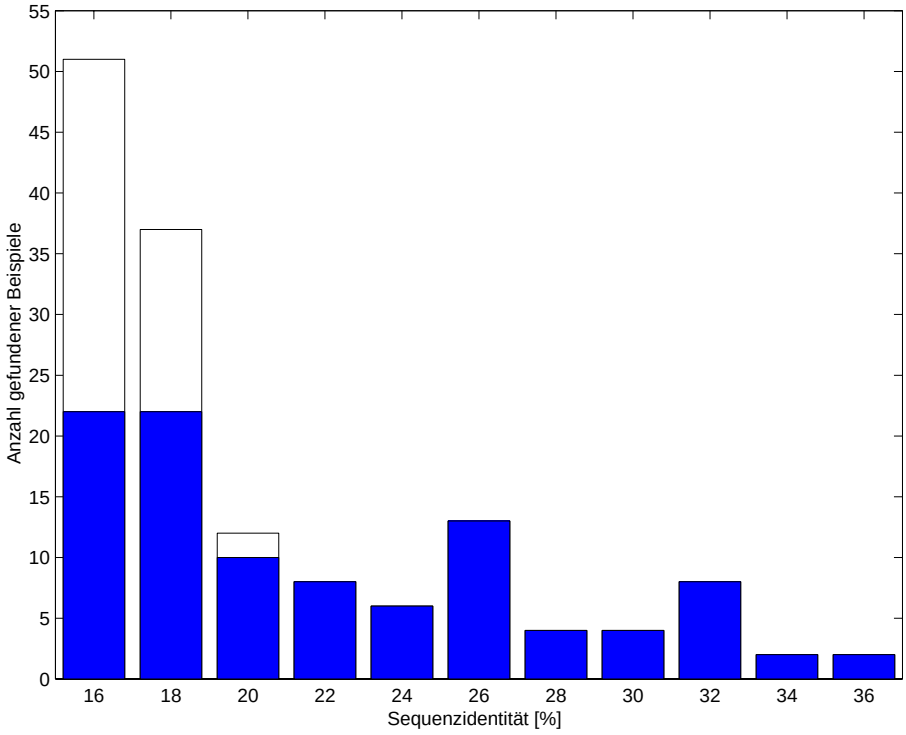


Abbildung 3.5: Abhängigkeit der Anzahl signifikanter (gefüllte Balken) und nicht signifikanter (offene Balken) struktureller Ähnlichkeiten von der Sequenzidentität in Prozent gefunden in einer repräsentativen Proteinmenge [7](siehe Text).

die Bewertung mindestens drei Standardabweichungen über dem Mittelwert der Verteilung der Ähnlichkeitsbewertungen zu anderen Proteinen liegt. Die aus [7] entnommene Abbildung 3.5 zeigt den Anteil der gemäß dieser Definition signifikanten strukturellen Ähnlichkeiten in Abhängigkeit von der Sequenzidentität. Die offenen Balken zeigen die Anzahl aller gefundenen strukturellen Ähnlichkeiten und die geschlossenen Balken die Ähnlichkeiten, die nach dem beschriebenen Kriterium als signifikant bewertet werden.
Abbildung 3.5 zeigt, daß strukturelle Ähnlichkeiten, die mit einer Sequenzidentität größer als ungefähr 20% verbunden sind, in der Regel als signifikant anzusehen sind. In der *twilight zone* sinkt die Anzahl signifikanter struktureller Ähnlichkeiten mit der Sequenzidentität.
Für die Proteinstrukturvorhersage in der *twilight zone* bedeutet diese systematische Analyse, daß signifikante strukturelle Ähnlichkeiten auch weit unterhalb der 25%–Identitätsmarke gefunden werden können und damit die Suche nach derartigen Ähnlichkeiten mit Verfahren sinnvoll ist, die eine Sequenz direkt gegen eine Struktur alinieren. Auf der anderen Seite zeigt dieser systematische strukturelle Vergleich auch, daß der Anteil der signifikanten Ähnlichkeiten von den mit bei al-

len auf derartige Weise postulierten Ähnlichkeiten mit der Sequenzidentität sinkt. Eine Vorhersage ist damit mit umso mehr Vorsicht zu behandeln, je geringer die Sequenzidentität ist.
Die helikalen Zytokine stellen hier in zweifacher Hinsicht ein gutes Beispiel für den Nutzen des strukturellen Vergleiches und der ähnlichkeitsbasierte Proteinstrukturvorhersage in der *twilight zone* dar. Ihre gemeinsame Abstammung ist nicht anhand der Sequenzen sondern nur auf Basis eines Strukturvergleichs erkennbar [293]. Außerdem wurde in einer echten Vorhersage unter Anwendung von *Threading*–methoden [211] für ein Protein, das mit der Fettleibigkeit in Zusammenhang gebracht wird, eine strukturelle Verwandtschaft zu den helikalen Zytokinen vorhergesagt, die mittlerweile durch Aufklärung der Struktur mit Hilfe der Röntgenkristallographie bestätigt wurde [375]. Dies zeigt, daß Proteinstrukturvorhersagemethoden in Abwesenheit von signifikanter Sequenzidentitäten auch zum Erkenntnisgewinn bei der Behandlung von Krankheiten beitragen können.

3.3.4 Strukturelle Ähnlichkeit und Funktion

Die Funktion eines Proteins kann nicht allein aus der Kenntnis des Faltungstyps vorhergesagt werden. So haben zum Beispiel viele Oxidasen, Isomerasen, Kinasen, Aldolasen, Synthasen, Dehydrogenasen und Proteasen entweder eine *TIM–barrel*–artige Faltung oder eine offene Faltblattstruktur [36]. Dies bedeutet, daß auf hohem Detailgrad betrachtet verschiedene Funktionen in der Natur durch sehr ähnliche Faltungsmuster realisiert sind und somit eine Funktionsvorhersage rein auf Basis des Faltungstyps nicht möglich ist. Die Spezifizität und Funktion eines Enzyms wird dagegen durch die Gestalt und Zusammensetzung des aktiven Zentrums des Proteins bestimmt.
Eine Funktionsvorhersage aus der Kenntnis der Faltung ist in vielen Fällen dennoch möglich, da die katalytisch aktiven Zentren aller Enzyme einer Faltungsklasse nahezu immer in der gleichen Strukturregion liegen. In der Regel wird das aktive Zentrum eines Enzyms durch die Schleifenbereiche zwischen den den Faltungstyp bestimmenden Sekundärstrukturelementen gebildet. Bei den *TIM–barrel*–artigen Strukturen sind dies die Schleifen zwischen den sich in der sequentiellen Reihenfolge abwechselnden Helices und Strängen an einem Ende des Fasses.
Ist es also neben der Vorhersage der Faltungsklasse auch möglich, ein korrektes und detailliertes Strukturmodell abzuleiten, so kann durch eine genauere Analyse des Strukturbereiches, der in dieser Faltungsklasse typischerweise das katalytisch aktive Zentrum ausmacht, die Funktion eines Proteins vorhergesagt werden. Dies ist eine Art der Funktionsvorhersage, wie sie bereits heute an experimentell aufgeklärten Proteinstrukturen durch strukturellen Vergleich durchgeführt wird. Von besonderem Interesse ist diese Art der strukturbasierten Funktionsvorhersage in Fällen, wo eine funktionelle Verwandtschaft oder Analogie nicht auf Sequenzebene erkannt werden kann sondern erst durch den strukturellen Vergleich zu

Tage tritt [151].

Leider impliziert die strukturelle Ähnlichkeit zweier Proteine nicht immer automatisch auch eine funktionelle Ähnlichkeit. So sind heute auch zahlreiche Beispiele bekannt, wo Proteine mit ähnlicher Struktur total unterschiedliche Funktionen haben [149]. Der Zusammenhang zwischen Struktur und Funktion ist sehr komplex, weshalb auch für die Struktur–Funktions–Beziehung in Analogie zur Sequenz–Struktur–Beziehung von einer *twilight zone* gesprochen wird [297].

Die beobachtete klare Trennung der für die Stabilität der Faltung verantwortlichen Aminosäuren im Kernbereich des Proteins von den für die spezifische Funktion des jeweiligen Enzyms notwendigen Aminosäuren in den Schleifenbereichen legt außerdem die Vermutung nahe, daß die Optimierung der Funktion und die Stabilisierung der für diesen Zweck verwendeten Faltung auch in der Evolution zwei unabhängige und parallel verlaufende Entwicklungen darstellen [36].

Für die theoretische Proteinstrukturvorhersage – zumindest in der Form, wie sie heute möglich ist und durchgeführt wird – bedeutet diese Tatsache einen großen Glücksfall. Würde jedes Enzym zur Erfüllung seiner Funktion und zur Ausbildung der zugehörigen Spezifizität einen grundlegend unterschiedlichen Faltungstyp ausbilden, wäre eine ähnlichkeitsbasierte Proteinstrukturvorhersage, wie sie auch Bestandteil dieser Arbeit ist, nicht möglich, da es dann keine sinnhafte Definition eines Ähnlichkeitsbegriffes geben würde.

3.3.5 Klassifizierung von Proteinen

Die Anzahl der strukturbekannten Proteine erscheint zwar hinsichtlich der Anzahl der bekannten Proteinsequenzen verhältnismäßig gering, jedoch ist es dem, der an den evolutionären Beziehungen zwischen Proteinen und den generellen Faltungsprinzipien interessiert ist, nicht mehr möglich, jede Proteinstruktur getrennt zu betrachten. Die obigen Betrachtungen bieten die Möglichkeit, Proteine nach verschiedenen Kriterien zu Gruppen zusammenzufassen und zu klassifizieren. Die Anwendung verschiedener Vergleichskriterien führt zu hierarchischen Klassifizierungen der bisher bekannten Strukturen.

In den vergangenen Jahren wurden in der Literatur verschiedene Klassifizierungsschemata für Proteinstrukturen diskutiert, von denen einige in Klassifizierungsdatenbanken umgesetzt wurden [89, 243, 260, 328]. Die zwei bekanntesten Klassifizierungen sind die von Orengo *et al.* generierte `CATH`–Klassifizierung [228, 260] und die von Murzin *et al.* abgeleitete `SCOP`–Klassifizierung [243]. Während `CATH` mit dem Ziel der weitgehend automatischen Klassifizierung der Proteine konzipiert wurde, werden bei `SCOP` automatische Vergleichsmethoden nur zur Unterstützung der weitgehend manuellen Klassifizierung eingesetzt. Dies läßt neben den teilweise unterschiedlichen Mengen der bereits klassifizierten Proteinen den Einsatz beider Klassifizierungsschemata zur Faltungserkennung sinnvoll erscheinen.

3.3.5.1 CATH

Ziel der Autoren von CATH (**C***lass* **A***rchitecture* **T***opology* **H***omology*) ist die automatische Klassifizierung von Proteinstrukturen in einem hierarchischen System in Analogie zu der allgemein verbreiteten Klassifizierung für Enzyme durch die sogenannten EC–Nummern [18, 31].
Im ersten Schritt werden alle Multidomän–Proteine in ihre Domänen aufgeteilt. Die automatische Bestimmung von Domängrenzen ist bislang nicht zufriedenstellen gelöst, obwohl es einige Lösungsansätze gibt [160, 170, 317, 334, 335], die in vielen Fällen brauchbare Lösungen erzeugen. Daher erfolgt die Bestimmung der Domängrenzen bei der CATH–Klassifizierung durch eine Konsensusbildung über die Ergebnisse dreier verschiedener Algorithmen [262]. Der nächste Schritt besteht in der Zusammenfassung der Proteine zu Familien, deren Sequenzen zu mehr als 35% identisch sind. Dazu werden Proteinsequenzen mittels der *single linkage cluster*–Methode [163] so zusammengefaßt, daß zwischen je zwei Elementen einer Familie eine Kette von paarweisen Alignments mit Sequenzidentität größer 35% existiert. Im nächsten Schritt werden aus den Familien ausgewählte Repräsentanten auf Strukturebene mit dem SSAP–Algorithmus [263, 338, 339] verglichen und auf Basis eines besonderen Vergleichswertes zusammengefaßt, falls dieser Vergleichswert einen bestimmten Grenzwert überschreitet. Strukturpaare, die diesen Grenzwert knapp unterschreiten, werden als analoge Strukturen [260] bezeichnet, da für sie nur eine schwache strukturelle Verwandtschaft erkannt wurde, aber eine evolutionäre Beziehung zwischen ihnen möglich ist. Da der SSAP–Algorithmus beim Vergleich die sequentielle Reihenfolge der Aminosäurereste beibehält, verfügen alle Proteine einer Familie über die gleiche topologische Anordnung der Sekundärstrukturelemente. Die Repräsentanten der so entstandenen Strukturfamilien und nicht automatisch eingeordnete Strukturen werden mittels Computergraphik entsprechend ihrer Architektur und Sekundärstrukturzusammensetzung klassifiziert.
Im Unterschied zur ersten Version [260] werden in der aktuellen Version von CATH [228, 262] nicht mehr Proteine sondern, wie in SCOP, Domänen als Basiseinheit für die Klassifizierung verwandt.
Aus dieser Vorgehensweise ergibt sich das folgende hierarchische Klassifizierungsschema:

- Die **Faltungsklasse C** wird bestimmt durch die Sekundärstrukturzusammensetzung:

  α : Der Anteil helikaler Sekundärstrukturelemente ist größer als 60%, der Anteil der β–Stränge kleiner als 5%.

  β : Mehr als 50% der Sekundärstrukturelemente sind β–Stränge und der Anteile der Helices liegt unter 5%.

  $\alpha\beta$: Die Struktur verfügt sowohl über Helices ($15 - 55\%$) als auch über Stränge ($10 - 45\%$).

Domänen, die keine oder sehr wenige regelmäßige Sekundärstrukturelemente, werden gesondert betrachtet und zu einer eigenen Klasse zusammengefaßt.

- Die **Architektur A** beschreibt die räumliche Anordnung der verschiedenen Sekundärstrukturelemente zueinander zu sogenannten Supersekundärstrukturelementen, wie zum Beispiel gebündelte Helices, einzelne Faltblätter, übereinanderliegende oder Faß–ähnliche Faltblattstrukturen.

- Die **Topologie T** zieht zusätzlich zur räumlichen Anordnung die sequentielle Reihenfolge der Sekundärstrukturelemente in Betracht, d.h. hier wird unterschieden, ob Stränge parallel oder antiparallel verlaufen. Domänen mit der gleichen Topologie sind superpositionierbar, da die Voraussetzung für diese Zusammenfassung eine gute SSAP–Bewertung ist. Eine ausführliche Beschreibung dieser häufig in Proteinen auftretenden Faltungsmotive findet sich in [40].

- Die **Homologie H** ist ein Kriterium, das die Proteine zusammenfaßt, deren evolutionäre Verwandtschaft aufgrund sehr ähnlicher Strukturen und funktioneller Ähnlichkeiten postuliert werden kann. Die so entstehenden Familien werden auch Superfamilien genannt.

- Die **Sequenzfamilienzugehörigkeit S** ist das einfachste Kriterium zum Vergleich von Proteinen. Hier werden die Proteine zu einer Familie zusammengefaßt, deren evolutionäre Verwandtschaft schon durch einen Vergleich ihrer Sequenzen eindeutig belegt werden kann.

Eine klare Trennung zwischen Proteinen, in denen Helices und Stränge vermischt auftreten ($\alpha + \beta$), und Proteinen, in denen sie alternierend auftreten (α/β), wie sie in der ersten Version von `CATH` [260] vorgenommen wurde, ist nicht immer möglich. Daher sind in der aktuellen Version von `CATH` [228] die Klassen ($\alpha + \beta$) und (α/β) zur Klasse ($\alpha\beta$) zusammengefaßt worden. Das `PDB`–Release vom September 1996 besteht entsprechend dieser Klassifizierung aus 5993 Proteinketten, 8078 Domänen, 1068 Sequenzfamilien, 645 homologen Superfamilien 505 Faltungsfamilien und 28 unterschiedlichen Architekturen [262].

3.3.5.2 SCOP

Im Gegensatz zur `CATH`–Klassifizierung haben Murzin *et al.* in `SCOP` (**S***tructural* **C***lassification* **o***f* **P***roteins*) [243] wesentlich durch die visuelle Inspektion und den visuellen Vergleich der Strukturen klassifiziert. Automatische Werkzeuge zum Sequenz– beziehungsweise Strukturvergleich wurden nur eingesetzt, um eine derartige Analyse auf zum damaligen Zeitpunkt etwa 3200 Proteinen mit vertretbarem Aufwand anwenden zu können. Ein weiterer Unterschied zu `CATH` besteht darin, daß `SCOP` bereits in seiner ersten Version als Klassifizierungseinheit die

einzelnen Domänen der Proteine verwendete. Wie CATH ist SCOP hierarchisch aufgebaut und klassifiziert die Domänen nach folgenden Kriterien:

- Die **Faltungsklasse** wird festgelegt durch die in der jeweiligen Domäne befindlichen Sekundärstrukturelemente:

  α : Die bestimmenden Strukturelemente sind α–Helices.

  β : Die bestimmenden Strukturelemente sind β–Stränge.

  $\alpha \,\&\, \beta$: Helices und Stränge sind stark miteinander vermischt.

  $\alpha + \beta$: Helices und Stränge sind klar getrennt.

  Proteine mit mehreren Domänen, theoretische Modelle und entworfene Proteine werden gesondert betrachtet.

- Eine gemeinsame **Faltung** liegt dann vor, wenn zwei Proteine nicht nur in den wesentlichen Strukturelementen sondern auch in deren räumlicher und topologischer Anordnung übereinstimmen. Domänen mit der gleichen Faltung können sich jedoch in den Schleifenbereichen grundlegend unterscheiden oder sogar zusätzliche Sekundärstrukturelemente besitzen. Allgemein ist ein gemeinsamer evolutionärer Ursprung für Proteine der gleichen Faltung denkbar, doch auf Grundlage des jetzigen Datenbestandes nicht in allen Fällen nachweisbar.

- In **Superfamilien** sind Domänen zusammengefaßt, deren strukturelle und in vielen Fällen auch funktionelle Eigenschaften einen gemeinsamen evolutionären Ursprung nahelegen. Dieser gemeinsame Ursprung ist jedoch nicht unbedingt schon auf Sequenzebene eindeutig nachweisbar.

- **Familien** bestehen aus Proteinen, die entweder untereinander eine Sequenzidentität größer als 30% haben oder aber so ähnliche Strukturen und Funktionen haben, daß ein gemeinsamer evolutionärer Ursprung offensichtlich ist.

Gemäß diesen Kriterien wurden die 3179 Proteindomänen, die in der Proteinstrukturdatenbank 1994 enthalten waren, zu 498 Familien, diese zu 366 Superfamilien und dieses wiederum zu 274 Faltungen zusammengefaßt. Von den 129 im Jahr 1995 veröffentlichten Strukturen konnten 57 bereits vorhandenen Superfamilien und weitere 29 Strukturen existierenden Faltungsklassen zugewiesen werden. Nur 43 Proteinstrukturen konnten keiner der bekannten Faltungsklassen zugeordnet werden [242].

3.3.6 Bedeutung für die Vorhersage

Die Bedeutung der obigen Betrachtungen für die Möglichkeiten zur theoretischen beziehungsweise computergestützten Proteinstrukturvorhersage läßt sich in folgenden Punkten zusammenfassen:

- Liegt eine hinreichend hohe Sequenzidentität zu einem oder gar mehreren strukturaufgeklärten Protein vor, so ist es möglich, für das untersuchte Protein ein zuverlässiges Strukturmodell auf Basis der bekannten Struktur zu generieren. Dies ist heute als Stand der Technik anzusehen, wenn auch bei weitem nicht alle Probleme zufriedenstellend gelöst sind, wie sich bei der späteren genaueren Betrachtung der Methoden 4.1 zeigen wird.
- Beispiele, bei denen die Verwandtschaft zweier Proteine nur auf der Strukturebene und nicht schon durch den Vergleich der Sequenzen erkennbar wird, geben Anlaß zu der Hoffnung, daß eine ähnlichkeitsbasierte Strukturvorhersage (siehe Definition 3.2) auch dann noch zum Erfolg führen kann.
- Klassifizierungsschemata wie `CATH` und `SCOP`, die strukturaufgeklärte Proteine entsprechend ihrer Struktur und auch Funktion klassifizieren, stellen insbesondere für die Methoden ein unerläßliches Hilfsmittel dar, die auf einer Erkennung von Faltungen basieren.
- Die Verteilung der bisher aufgeklärten Proteinstrukturen auf die verschiedenen Faltungsklassen läßt vermuten, daß in der Natur bestimmte Faltungstypen bevorzugt werden [262]. Es scheint sich dabei um besonders stabile Faltungsformen zu handeln. Daher besteht Grund zur Hoffnung, daß neue Sequenzen mit Faltungserkennungsmethoden diesen besonders stabilen Faltungstypen zugeordnet werden können.
- Das langsame Wachstum der Anzahl neuer Faltungsklassen im Verhältnis zur Anzahl neu aufgeklärter Strukturen zeigt, daß zumindest auf dem Abstraktionsniveau der Faltungsklasse bereits eine zwar nicht repräsentative aber auch nicht eine auf Spezialfälle eingeschränkte Sicht aller in der Natur vorkommenden Proteinstrukturen existiert.
- In aktuellen Schätzungen [42] wird aufgrund der geringen Anzahl neu entdeckter Faltungstypen davon ausgegangen, daß mit dem Abschluß des *Human Genome Projects* die meisten der in der Natur vorkommenden Faltungstypen bekannt sein werden. Da die Auswahl der Proteine für die Strukturaufklärung mit der Maßgabe der Kristallisierbarkeit und nicht mit dem Ziel der vollständigen Überdeckung des Strukturraumes erfolgt, wird es jedoch weit mehr Zeit erfordern, bis alle Faltungstypen bekannt sind.

Daraus folgt, daß Verfahren zur Faltungserkennung für die Lösung strukturspezifischer Fragstellungen in den nächsten Jahren eine sehr große Bedeutung zukommen wird. Und solange nicht auch Repräsentanten aller Superfamilien strukturaufgeklärt sind und damit Verfahren zur Aufdeckung entfernter Sequenzverwandtschaften eingesetzt werden können, stellen Faltungserkennungsmethoden neben der aufwendigen experimentellen Strukturaufklärung den einzigen Weg zur Ableitung strukturbezogener Informationen zu einer Vielzahl von Proteinen dar.

Kapitel 4

Methoden zur Proteinstrukturvorhersage

Heute ist es mit Erfolg möglich, Proteinstrukturen beziehungsweise partielle Eigenschaften von Proteinstrukturen mit computergestützten Methoden vorherzusagen. Dies liegt daran, daß – wie bereits erwähnt – verwandte Proteine auch strukturell Gemeinsamkeiten haben und im nativen Zustand ähnliche Faltungsstrukturen annehmen. Diese Erkenntnis wird von den im folgenden beschriebenen Methoden zur theoretischen Proteinstrukturvorhersage ausgenutzt. Dabei wird von dem Begriff der Energie einer Konformation weitestgehend abstrahiert und versucht, entweder diese Ähnlichkeiten zu erkennen und in die Vorhersage einfließen zu lassen oder die Menge der bekannten Strukturen zur Ableitung und Kalibrierung von empirischen Kraftfeldern zu nutzen. Üblicherweise werden die theoretischen Strukturvorhersagemethoden grob wie folgt eingeteilt [90]:

Vergleichende Modellierung: Die Methode der vergleichenden Modellierung wird erfolgreich angewendet in Fällen, wo die Sequenzidentität zwischen dem strukturbekannten Protein und der zu untersuchenden Sequenz größer als 50% ist. In diesen Fällen kann sowohl die Auswahl der verwandten Struktur als auch die Abbildung der Sequenz auf die Struktur mit Sequenzalignmentmethoden erfolgen. Mit sinkender Sequenzidentität steigt die Häufigkeit von Fehlern, die bei der Abbildung der Sequenz auf die Struktur gemacht werden.

Faltungserkennung: Methoden zur Faltungserkennung versuchen, Homologien zu bekannten Strukturen aufzudecken, wenn das mit rein die Sequenzähnlichkeit bewertenden Kostenfunktionen und Methoden nicht mehr möglich ist. Zur Berechnung der Abbildung der Sequenz auf die Struktur werden dabei sogenannte Sequenzstrukturalignment– oder *Threading*–Methoden eingesetzt, die im Bereich von Sequenzidentitäten unter 50% zuverlässigere Alignments berechnen als reine Sequenzalignmentmethoden und in der Regel auch noch unter 25% Sequenzidentität brauchbare Modelle liefern.

***Ab initio*–Methoden:** Methoden, die nicht oder nur indirekt auf die Homologie zu bereits bekannten Strukturen aufbauen, werden als *Ab initio*–Methoden bezeichnet.

Inverse Faltung: Methoden zur inversen Faltung starten mit dem Modell einer dreidimensionalen Struktur und haben den Entwurf einer Proteinsequenz zum Ziel, die mit hoher Wahrscheinlichkeit und Stabilität unter natürlichen Bedingungen diese vorgegebene Faltung annimmt [41, 166, 372].

Der zielgerichtete Entwurf von Proteinen, wie er zum Beispiel Ziel der Methoden zur Lösung des inversen Faltungsproblems ist, hat gegenwärtig sowohl unter Forschungs– als auch unter Anwendungsgesichtspunkten nur geringe Bedeutung, da zu wenig über die der Proteinfaltung zugrundeliegenden Regeln bekannt ist. Das Design von Proteinen beschränkt sich gegenwärtig auf die zielgerichtete Modifikation natürlicher Proteine durch Punktmutationen einzelner Aminosäuren in den aktiven Zentren [337] oder zur Stabilisierung von Faltungen [145] zum Beispiel durch die Ausbildung zusätzlicher Salzbrücken [11].
Die größte Bedeutung kommt den Verfahren zur Vorhersage struktureller Eigenschaften von aus der Natur bekannten Proteinsequenzen zu. Daher werden die zur Vorhersage verwandten Verfahren im folgenden anhand der in der Literatur vorgeschlagenen Methoden diskutiert.

4.1 Vergleichende Modellierung

Wie bereits erwähnt, kommt die Methode der vergleichenden Modellierung dann zum Einsatz, wenn ein oder mehrere strukturaufgeklärte Proteine eine signifikante Sequenzhomologie zu der untersuchten Sequenz aufweisen. Das Finden der homologen Strukturen ist in der Regel schon mit recht einfachen Methoden, wie zum Beispiel `BLAST` [10] und `FASTA` [207, 274], möglich. Sei also im folgenden vorausgesetzt, daß bereits eine oder mehrere verwandte Strukturen identifiziert wurden.
Die Modellierung einer Proteinstruktur auf Basis der bekannten homologen Struktur – im folgenden auch Strukturvorlage genannt – erfolgt in der Regel in folgenden Schritten:

1. **Alignment**: Berechnung eines Alignments der Sequenz des untersuchten Proteins mit der Sequenz der bekannten Struktur.

2. **Rückgratplazierung**: Übertragung der Raumkoordinaten der Rückgratatome der Struktur auf die alinierten Reste der Sequenz gemäß des zuvor berechneten Alignments.

3. **Schleifenmodellierung**: Berechnung von Raumkoordinaten für die in Phase 2 nicht plazierten Rückgratatome, die zu nicht alinierten Resten gehören.

4. **Seitenkettenplazierung**: Plazierung der Seitenkettenatome entlang des modellierten Rückgrates.

5. **Nachoptimierung**: Optimierung der berechneten Struktur mit Energieminimierungs– bzw. Molekulardynamikmethoden.

Im folgenden wird die Vorgehensweise bei der vergleichenden Modellierung anhand zweier sogenannter *TIM–barrel*–strukturen, *Triose Phosphate Isomerasen*

vom Geißeltierchen (PDB: 1tpe) und vom Huhn (PDB:1tim), beispielhaft beschrieben. Zur Vereinfachung der Darstellung handelt es sich bei diesen zwei Beispielen um Proteine, deren Strukturen bereits mit experimentellen Methoden aufgeklärt wurden. Abbildung 4.1 zeigt ein Sequenzalignment der Proteinsequenzen. Da die

```
Alignment 1   Profile = (PDB1tpe+PDB1timA)

PDB1tpe          _000:MSKPQPIAAANWKCNGSQQSLSELIDLFNSTSINHDVQCVVASTFVHLAMTKERLSHPKFVIA
PDB1timA         _000:-APRKFFVGGNWKMNGKRKSLGELIHTLDGAKLSADTEVVCGAPSIYLDFARQKLDAKIGVAA
score            _000:          NWK NG   SL ELI          D   V       L      L     V A
reliability      _000: 000000000135578666887753111111111111111111111111111111100000000
SECSTR 'PDB1tpe' _000:      eeeee      hhhhhhhhhhhh        eeeee     hhhhhh     eeee
SECSTR 'PDB1timA'_000:      eeeee      hhhhhhhhhhhhh       eeeee    hhhhhhhh    eeeee

PDB1tpe          _060:AQNAIAKSGAFTGEVSLPILKDFGVNWIVLGHSERRAYYGETNEIVADKVAAAVASGFMVIAC
PDB1timA         _060:QNCYKVPKGAFTGEISPAMIKDIGAAWVILGHSERRHVFGESDELIGQKVAHALAEGLGVIAC
score            _060:        GAFTGE S    KD G  W  LGHSERR   GE  E    KVA A A G  VIAC
reliability      _060:000000002468986644446886664468999886668866644446866668776689999
SECSTR 'PDB1tpe' _060:e              ehhhhhhh         hhhhhh    hhhhhhhhhhhhh  eeeee
SECSTR 'PDB1timA'_060:e               hhhhhhh         hhhhhhh   hhhhhhhhhhhhh    eeee

PDB1tpe          _120:IGETLQERESGRTAVVVLTQIAAIAKKLKKADWAKVVIAYEPVWAIGTGKVATPQQAQEAHAL
PDB1timA         _120:IGEKLDEREAGITEKVVFQETKAIADNVK--DWSKVVLAYEPVWAIGTGKTATPQQAQEVHEK
score            _120:IGE L ERE G T  VV     AIA   K  DW KVV AYEPVWAIGTGK ATPQQAQE H
reliability      _120:998888898888866664444445531111  133578889999999999888998899866644
SECSTR 'PDB1tpe' _120:e   hhhhh   hhhhhhhhhhhhh    hhhhhheeeee             hhhhhhhhhh
SECSTR 'PDB1timA'_120:eee hhhhhh  hhhhhhhhhhhhhhh        eeeeeeehhh         hhhhhhhhhh

PDB1tpe          _180:IRSWVSSKIGADVRGELRILYGGSVNGKNARTLYQQRDVNGFLVGGASLKPEFVDIIKATQ
PDB1timA         _180:LRGWLKTHVSDAVAVQSRIIYGGSVTGGNCKELASQHDVDGFLVGGASLKPEFVDIINAKH
score            _180: R W        V    RI YGGSV G N   L  Q DV GFLVGGASLKPEFVDII A
reliability      _180:46664444444466666766888986668666866668888999889999999887766
SECSTR 'PDB1tpe' _180:hhhhhhhh  hhhhhh             hhhhh                 hhhhhh
SECSTR 'PDB1timA'_180:hhhhhhhhh hhhhhh           hhhhhh                 hhhhhh

Alignment                      value   PerPos.    Prof-1    Prof-2
Alignment length          =      250                250       247
Alignment ids             =      122     48.80 %   48.80 %   49.39 %
Alignment aligned         =      247     98.80 %   98.80 %  100.00 %
Alignment gaps            =        2      0.80 %    0.80 %    0.81 %
Alignment indels          =        3      1.20 %    1.20 %    1.21 %
```

Abbildung 4.1: Paarweises Sequenzalignment zweier *TIM–barrel*–strukturen (1tpe und 1timA): Die Sequenzidentität beträgt fast 50%, die Sekundärstrukturelemente sind gut aufeinander abbildbar (vergleiche Annotationszeilen SECSTR).

Sequenzidentität der zugehörigen Sequenzen nahe 50% liegt, stellt das Paar – nimmt man eine Struktur zum Beispiel die von 1tim Kette A als unbekannt an – ein typisches Beispiel für die Anwendung der vergleichenden Modellierung dar. Aus der Rückgratplazierung resultiert ein erstes Modell, das vor allem die Lage der *Core*–Bereiche durch Übernahme aus der Vorlagestruktur, (dem sogenannten *template*), festlegt. Diese sind in dem linken Teil der Abbildung 4.2 farbig hervorgehoben. Der rechte Teil der Abbildung 4.2 zeigt ein erstes Rückgratmodell, bei dem die in der Schleifenmodellierung hinzugekommenen Schleifenbereiche farbig markiert sind.
In der Regel ist es sinnvoll, nicht nur die Koordinaten der Rückgratatome der alinierten Positionen, sondern auch die Konformationen der konservierten Seitenketten zu übernehmen (linker Teil der Abbildung 4.3). Das Strukturmodell wird vervollständigt, indem die noch fehlenden Seitenketten hinzu modelliert werden

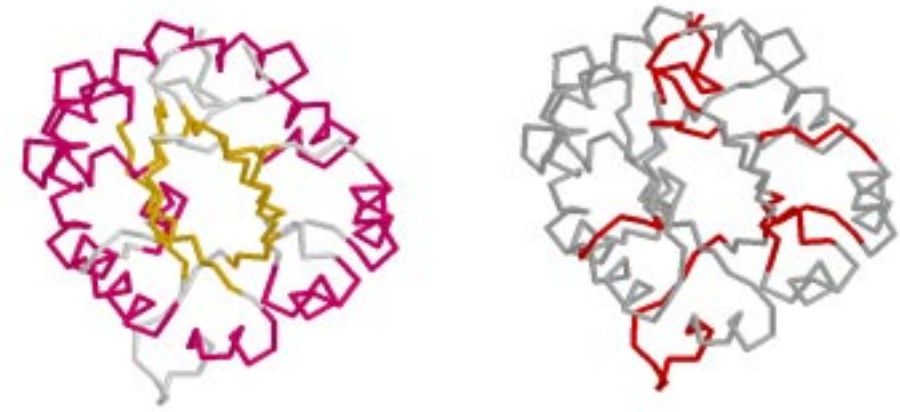


Abbildung 4.2: Vergleichende Modellierung: Modell nach Rückgratplazierung (links) und Schleifenmodellierung (rechts) für das Protein `1timA`.

(im rechten Teil der Abbildung 4.3 farbig hervorgehoben).

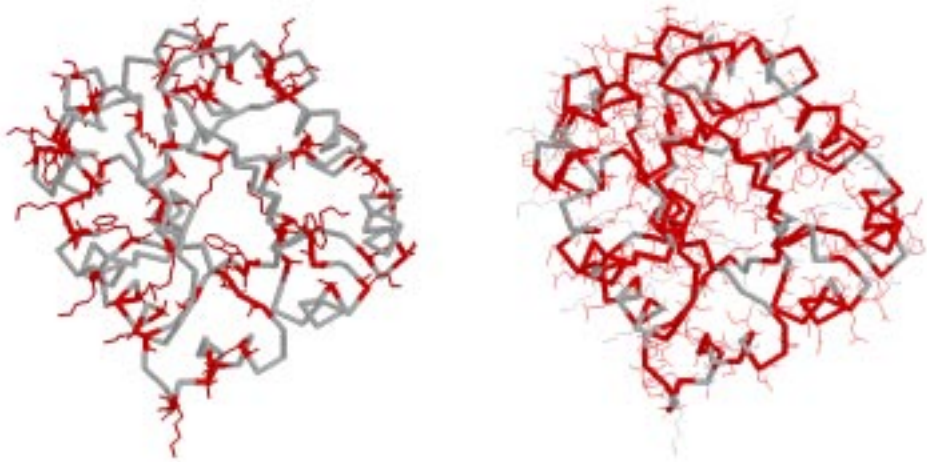


Abbildung 4.3: Vergleichende Modellierung: Modell nach der Übernahme der Seitenkettenkonformationen der konservierten Seitenketten (links) und nach Modellierung der restlichen Seitenketten (rechts) für das Protein `1timA`.

Da das grobe Gerüst des zu modellierenden Proteins von dem in der Strukturdatenbank gefundenen homologen Protein übernommen wird, stellt die Berechnung der Abbildung der Sequenz in die Struktur, die in der vergleichenden Modellierung durch Sequenzalignmentmethoden erfolgt, den entscheidenden Schritt dar. Zur Erkennung einer Homologie sind häufig schon lokal begrenzte Sequenzähnlichkeiten hinreichend. Um jedoch mit Sequenzalignmentmethoden eine zuverlässige, strukturell und damit biologisch richtigen Abbildung der untersuchten Sequenz

in die Struktur vorschlagen zu können, bedarf es zudem einer Sequenzähnlichkeit, die über alle Teile der Sequenz verteilt ist, da sonst Fehler bei der Abbildung unvermeidbar sind. Fehler in dieser Abbildung bedeuten aber immer auch, daß falsche Koordinaten für die an diesen Fehlern beteiligten Aminosäurereste in das Modell übernommen werden. Schon die Ergebnisauswertung [237] des ersten Workshop zum *Critical Assessment of Methods of Protein Structure Prediction* `CASP I` hat deutlich gezeigt, daß keine der bekannten Methoden zur vergleichenden Modellierung in der Lage ist, diese Fehler wieder auszumerzen.
Daher sind Sequenzalignmentmethoden, obwohl das algorithmische Problem weitgehend gelöst ist, immer noch Gegenstand der aktuellen Forschung und werden im Abschnitt 4.1.2 ausführlicher beschrieben, während die anderen Aspekte der vergleichenden Modellierung nur kurz angerissen werden sollen.
Die in dieser Dissertation entwickelte Methode der **R**ekursiven **D**ynamischen **P**rogrammierung (RDP) hat unter anderem auch die Verbesserung der Abbildungen von Sequenzen auf Strukturen zum Ziel, wo die Sequenzidentität unter 50% absinkt und sich bei Verwendung rein sequenzorientierter Verfahren erste Fehler im Alignment einschleichen, die sich teilweise als grobe Fehler in den daraus erzeugten Modellen auswirken (siehe dazu `CASP I` [237]).
Ein weiteres Ziel der RDP–Methode ist es, die Methode der ähnlichkeitsbasierten Proteinstrukturvorhersage auf Proteine geringer Sequenzhomologie auszudehnen und damit strukturelle Beziehungen, wie sie durch den Strukturvergleich aufgedeckt werden [7], vorherzusagen. Für diese Anwendungsfälle kommt es besonders darauf an, auch für Beispiele geringer Sequenzhomologie oder für Beispiele, wo sich die Sequenzähnlichkeit auf lokale Bereiche beschränkt, zuverlässige Modellvorschläge beziehungsweise Sequenzstrukturalignments zu berechnen.

4.1.1 Verschiedene Ansätze zur vergleichenden Modellierung

Die verschiedenen Ansätze zur vergleichenden Modellierung unterscheiden sich im wesentlichen in der Art und Weise, wie aus dem Alignment das Strukturmodell generiert wird. In einem der ersten Ansätze zur vergleichenden Modellierung [33] werden zunächst Sequenzalignments zu den verschiedenen in der Datenbank vorhandenen ähnlichen Proteinen berechnet. In diesen verschiedenen Alignments werden die jeweils ähnlichsten Sequenzabschnitte und die dazugehörigen Strukturfragmente bestimmt. Diese Strukturfragmente werden dann starr in das Rahmengerüst eingebaut, das aus den C_α–Koordinaten der in den Faltungen konservierten Bereiche gemittelt wurde, um so ein atomares Modell für das untersuchte Protein zu erhalten [33]. Eine umfassende Beschreibung und Analyse dieser Vorgehensweise, insbesondere des zu dieser Kategorie gehörigen Progammpaketes `COMPOSER` [332] findet sich in [165].
Ein weiterer Ansatz besteht darin, die verschiedenen über Sequenzalignmentmethoden gefundenen strukturellen Vorlagen in Fragmente aufzubrechen und die so entstehenden Lücken mit in der Datenbank gefundenen gut passenden Fragmen-

ten aufzufüllen. In dem Werkzeug LOOK [204] erfolgt die Auswahl eines Fragments nach der Sequenzidentität, der konformationellen Ähnlichkeit und der Kompatibilität mit dem bisher gebildeten Strukturmodell.
Der gegenwärtig vielversprechendste und auch von den meisten aktuellen Methoden benutzte Ansatz besteht darin, ein Modell auf Basis der bereits identifizierten strukturellen Vorlagen zu erzeugen, das den räumlichen und weiteren Randbedingungen genügt, die aus Analyse und Vergleich bekannter Proteinstrukturen abgeleitet sind [306]. Zur Lösung kommen dabei entweder die Methoden der Distanzgeometrie zum Einsatz (wie zum Beispiel in DRAGON [16]), oder das Problem wird als Optimierungsproblem formuliert, das dann zum Beispiel in MODELLER [305] mit der Methode der konjugierten Gradienten gelöst wird.
Eine übersichtsartige Beschreibung und ein Vergleich der verschiedenen Ansätze finden sich in [303] und [309]. Ergänzt werden diese Verfahren, die auf die Vorhersage der vollständigen Proteinstruktur zielen, durch eine Reihe von Methoden, die nur auf die Lösung von Teilproblemen abheben, wie zum Beispiel die Modellierung von Schleifenbereichen oder die Vorhersage der Seitenkettenkonformationen. Die Modellierung von Schleifenbereichen erfolgt heute in der Regel durch Suche nach geeigneten Schleifen in der Strukturdatenbank [82, 95, 219, 258, 295, 349], durch Konformationssuche mit optional nachgeschalteter Energieminimierung [376] oder durch eine Kombination dieser Methoden.
Vásquez gibt eine aktuelle Übersicht zur Vorhersage von Seitenkettenkonformationen in [350]. Eine Abschätzung über die Genauigkeit der mit diesen Methoden erzeugten Seitenkettenplazierungen geben Chung und Subbiah in [62]. Diese Analyse zeigt, daß die *RMS*–Abweichung der modellierten Seitenkettenkonformationen zu den experimentell bestimmten Konformationen um die 1.5Å liegt, sofern die Sequenzidentität zwischen der Strukturvorlage und dem modellierten Protein größer als 40% ist. Fällt die Sequenzidentität, so nimmt auch die Genauigkeit der Seitenkettenmodellierung rapide ab und überschreitet sehr schnell die 3Å–Marke. Ein Grund ist dafür sicher auch der größere Fehler mit dem schon das Rückgratmodell auch infolge der in diesem Bereich nicht zuvernachlässigenden Unterschiede in den Schleifenbereichen behaftet ist, den die Autoren mit etwa 2Å für den Bereich zwischen 20% und 25% Sequenzidentität angeben.
Insgesamt läßt sich sagen, daß die Verfahren zur vergleichenden Modellierung einen Stand erreicht haben, der es erlaubt, diese Verfahren zur Untersuchung vieler praxisrelevanter Probleme einzusetzen. Viele der Methoden sind bereits in kommerziell vertriebene Softwareprodukte [309] integriert und werden in der pharmazeutischen Industrie in großem Umfang eingesetzt. Ihre Grenzen finden diese Methoden dort, wo entweder eine sehr hohe Modellgenauigkeit gefordert ist oder wo die Sequenzidentität zu einer bekannten Struktur sehr gering ist. So sind zum Beispiel bereits bei einer Sequenzidentität von etwas mehr als 30% in der Regel bereits 20% der Reste falsch aliniert [309], was bereits für das Rückgratmodell Abweichungen von mehreren Angstrœm zur Folge hat. In einigen Fällen können aber auch derart grobe Modelle erste Hinweise zur Erklärung von Wirkmecha-

nismen geben. Dies gilt insbesondere in den Fällen, wo ein strukturell hochkonserviertes Motiv, wie zum Beispiel eine ATP–Bindestelle, Bestandteil der aktiven Stelle eines Proteins ist [182], und sich somit für diese interessante Region ein wesentlich genaueres Strukturmodell als für den Rest des Proteins generieren läßt [59, 126, 381].

4.1.2 Sequenzalignment

Sequenzalignment ist die Methode zur Aufdeckung evolutionärer Verwandtschaften, wie sie bei der vergleichenden Modellierung ausgenutzt werden. Wenn bei der Berechnung des optimalen Alignments Einfügungen von Insertionen und Deletionen erlaubt sind, erhält man ein Optimierungsproblem, das wie folgt definiert werden kann:

Definition 4.1 (Optimierungsproblem paarweises Sequenzalignment) *Seien* $A =< a_1, \ldots, a_n >$ *und* $B =< b_1, \ldots, b_m >$ *zwei Folgen mit* $a_i, b_i \in \Sigma$*, wobei* Σ *das Alphabet der* 20 *proteinogenen Aminosäuren ist. Seien* $\alpha, \beta \in \mathbb{R}$ *und sim eine Funktion* $\Sigma \times \Sigma \longrightarrow \mathbb{R}$*. Dann ist der Wert eines Alignments* $Score_h(A, B)$ *wie folgt definiert:*

$$\begin{aligned} Score_h(A,B) &= \sum_{a_i \in M_h(A,B) \cup N_h(A,B)} sim(a_i, b_{h(i)}) \\ &+ \alpha \, |G_h(A,B)| \\ &+ \beta \, |D_h(A,B) \cup I_h(A,B)| \end{aligned}$$

Ein optimales Alignment $h_{opt} \in \mathcal{H}(A, B)$ *ist definiert durch folgende Gleichung:*

$$Score_{h_{opt}}(A,B) = \max_{h \in \mathcal{H}(A,B)} Score_h(A,B)$$

Die Funktion *sim* bewertet dabei die Ähnlichkeit zweier zueinander alinierter Aminosäuren beziehungsweise Aminosäurepositionen. Diese Bewertung erfolgt in der Regel anhand von 20×20–Ähnlichkeitsmatrizen (siehe dazu Abschnitt 5.1). Jede Deletion oder Insertion wird mit dem Wert α bestraft. Die Eröffnung eines Gaps kostet β. Die Ähnlichkeitsmaße und die Probleme mit der geeigneten Wahl der Parameter der Gapkostenfunktion werden in Abschnitt 5.1 thematisiert.
Glücklicherweise ist zur Lösung des obigen Optimierungsproblems nicht die Generierung aller möglichen Alignments zwischen den zwei betrachteten Sequenzen erforderlich. Da die Anzahl der möglichen Alignments zwischen zwei Sequenzen exponentiell in der Länge der Sequenzen ist, würde ein auf der Enummerierung aller Alignments basierendes Optimierungsverfahren exponentielle Laufzeit haben.

Aufgrund der Struktur des Problems kann zur Lösung des Optimierungsproblems die Methode der *dynamischen Programmierung* angewandt werden. Hier wird ausgenutzt, daß sich der Wert des optimalen Alignments rekursiv aus den Werten der optimalen Alignments aller Präfixe der betrachteten Sequenzen berechnen läßt. Basierend auf dieser Optimierungsstrategie wurde 1970 von Needleman und Wunsch ein erster Algorithmus zur Berechnung des optimalen Alignments zweier Sequenzen vorgeschlagen [250]. Der veröffentlichte Algorithmus bestraft nur die Eröffnung von Gaps (d.h. $\alpha = 0$) und hat eine Laufzeit kubisch in der Länge der Sequenzen. Das vorgeschlagene Optimierungsschema kann jedoch auf allgemeine Gapkostenfunktionen übertragen werden, ohne daß die Laufzeit asymptotisch zunimmt [226].
Für lineare Gapkosten kann die Lösung mit der Methode von Gotoh in Zeit $\mathcal{O}(nm)$ erfolgen [124]. Für konvexe beziehungsweise konkave Gapkostenfunktionen existieren implementierungsaufwendige Algorithmen, die die Laufzeit auf $\mathcal{O}(nm\log(min(n,m)))$ [107] beziehungsweise $\mathcal{O}(nm)$ [229, 362] reduzieren. Diese Varianten sind jedoch ohne große Bedeutung für praktische Anwendungen.
Allgemein werden jedoch – wie schon in Definition 4.1 angegeben – affine Gapkostenfunktionen verwendet [102]. Affine Gapkosten erlauben die Anwendung des quadratischen Algorithmus von Gotoh [124]. Für affine Gapkosten kann die Methode der dynamischen Programmierung durch folgende Rekursion beschrieben werden, die drei $n \times m$–Matrizen D, R und C mit Einträgen füllt:

$$D_{i,j} = \text{MAX}(D_{i-1,j-1} + sim(a_i, b_j), R_{i,j}, C_{i,j})$$

$$\begin{aligned} R_{i,j} &= \text{MAX}(D_{i,j-1} - \alpha - \beta, R_{i,j-1} - \alpha) \\ C_{i,j} &= \text{MAX}(D_{i-1,j} - \alpha - \beta, C_{i-1,j} - \alpha) \end{aligned}$$

Der Eintrag $D_{i,j}$ der Matrix D enthält also nach Abarbeitung der Positionen i und j den *Score* des optimalen Alignments der Präfixe $A_{|i}$ und $B_{|j}$ der Sequenzen A und B. Die Matrizen R und C enthalten den *Score* des optimalen Alignments der Präfixe $A_{|i}$ und $B_{|j}$ unter der Bedingung, daß die letzte Position im Alignment eine Insertion beziehungsweise Deletion war. Damit berechnen sich die Einträge dieser Matrizen auf eine der folgenden Arten:

1. Fortsetzung des Präfixalignments $h(A_{|i-1}, B_{|j-1})$ durch eine Identität oder Nichtübereinstimmung.

2. Fortsetzung des Präfixalignments $h(A_{|i}, B_{|j-1})$ durch Eröffnung eines neuen Gaps in der Sequenz A.

3. Fortsetzung des Präfixalignments $h(A_{|i}, B_{|j-1})$ durch Verlängerung eines Gaps in der Sequenz A.

4. Fortsetzung des Präfixalignments $h(A_{|i-1}, B_{|j})$ durch Eröffnung eines neuen Gaps in der Sequenz B.

5. Fortsetzung des Präfixalignments $h(A_{|i-1}, B_{|j})$ durch Verlängerung eines Gaps in der Sequenz B.

Nach Terminierung des Algorithmus enthält der Matrixeintrag $D_{n,m}$ den *Score* des optimalen Alignments der zwei Sequenzen. Das Auslesen des beziehungsweise der optimalen Alignments kann durch eine rekursive Prozedur erfolgen, die den Pfad oder die Pfade durch die Matrix zurückverfolgt, für die die Maximumsbildung mit Gleichheit erfüllt sind. Der Zeitaufwand pro ausgelesenem Alignment entspricht der Länge des Alignments, also $O(m+n)$.
Anhand des Berechnungsschemas wird deutlich, daß zur Berechnung des optimalen Alignments unter der Annahme affiner Gapkosten drei Matrizen der Dimension $n \times m$ benötigt werden. Der Platzbedarf des Algorithmus wächst also quadratisch in der Länge der Sequenzen. Für den Vergleich von Proteinsequenzen, deren Länge im allgemeinen kleiner als 1000 Aminosäurereste ist, spielt der Platzbedarf daher nur eine untergeordnete Rolle. Ist man nur an dem Wert des optimalen Alignments interessiert oder bereit, für das Ausgraben des Alignments ebenfalls quadratische Zeit zu investieren, kann der Speicherbedarf mit dem Trick von Hirschberg [147] auf lineare Größe eingeschränkt werden [245]. Dies ist jedoch nur in wenigen Fällen sinnvoll, zum Beispiel beim Vergleich genomischer Sequenzen auf Parallelrechnern, deren Knoten nur über wenig Hauptspeicher verfügen.
Nicht immer ist es sinnvoll, alle Gaps in einem Alignment uniform zu bewerten. Zum Beispiel ist es beim Vergleich von Teilstücken einer Sequenz mit vollständigen Sequenzen nicht angezeigt, die Gaps, die notwendig sind, um den Längenausgleich herzustellen, genauso zu bewerten, wie Gaps, die in sinnhaften Bereichen des Alignments auftreten. Überlegungen dieser Art haben zur Entwicklung verschiedener Alignmentmodi geführt:

global: Globale Alignments enthalten alle Positionen beider Sequenzen und notwendige Insertionen beziehungsweise Deletionen werden uniform bestraft.

***free–shift*:** Bei *free–shift*–Alignments ist es in beiden Sequenzen erlaubt, am Anfang und am Ende des Alignments Gaps einzuführen, ohne daß diese durch die Kostenfunktion bestraft werden.

***substring*$\star$:** Teilsequenzalignments sind asymmetrische Verallgemeinerungen des *free–shift*–Alignments in dem Sinne, daß in beiden Sequenzen Anfangs– und Endgaps bestraft oder nicht bestraft werden können. Es gibt also 4^2 unterschiedliche Alignmentmöglichkeiten. Werden Gaps an allen Enden (an keinem Ende) bestraft, erhält man ein globales (*free–shift*) Alignment. Der $\star$ steht hier für eine der 14 zusätzlichen Kombinationen.

Beim Alignment einer kürzeren Sequenz gegen eine wesentlich längere kann es zum Beispiel sinnvoll sein, Gaps am Anfang und Ende der längeren Sequenz ohne zusätzliche Kosten zu erlauben, aber gleichzeitig zu verlangen, daß die gesamte kürzere Sequenz vollständig aliniert wird.

local: Bei lokalen Alignments wird das beste Alignment der optimalen Alignments aller Teilzeichenketten der Sequenzen gesucht.

Abbildung 4.4 veranschaulicht schematisch sowohl die Unterschiede der verschiedenen Modi als auch den Berechnungsprozeß für affine Gapkosten. Um die Beziehung zwischen der Rekursionsgleichung und der Abbildung herzustellen, wurde hier die Matrixdarstellung [314] anstelle der von anderen Autoren bevorzugten Graphenrepräsentation [56, 244, 247] gewählt. Die Implementierung dieser un-

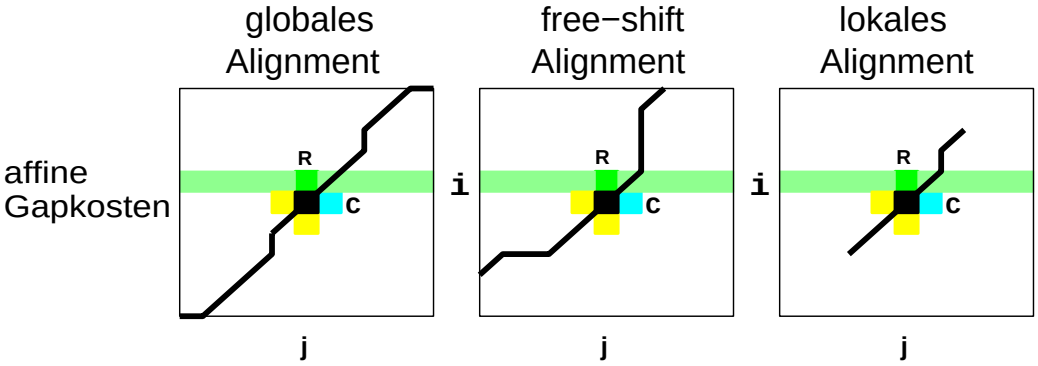


Abbildung 4.4: Schematische Darstellung der verschiedenen Alignmentmodi und ihrer Berechnung für affine Gapkostenfunktionen.

terschiedlichen Alignmentmodi kann durch das Rekursionsschema 4.1 erfolgen, indem die erste Spalte und Zeile der Matrizen unterschiedlich initialisiert werden [226]. Allein das lokale Alignment nimmt hier eine Sonderrolle ein, da es auf den ersten Blick die Berechnung von $n \times m$ optimalen Teilsequenzalignments erfordert. Doch Smith und Waterman [327] haben durch eine einfache aber wirksame Modifikation des obigen Rekursionsschemas folgende effiziente Lösung vorgeschlagen:

$$D_{i,j}^{local} = \text{MAX}(\quad \mathbf{0} \quad , D_{i-1,j-1}^{local} + sim(a_i, b_j), R_{i,j}, C_{i,j})$$

$$\begin{aligned} R_{i,j} &= \text{MAX}(D_{i,j-1}^{local} - \alpha - \beta, R_{i,j-1} - \alpha) \\ C_{i,j} &= \text{MAX}(D_{i-1,j}^{local} - \alpha - \beta, C_{i-1,j} - \alpha) \end{aligned}$$

Dabei wird ausgenutzt, daß Alignmentbereiche, die einen negativen *Score* haben, niemals zu einem optimalen lokalen Alignment gehören können. Der *Score* des optimalen lokalen Alignments kann somit durch Bestimmung des maximalen Eintrags in der Matrix D^{local} erfolgen. Damit können der *Score* und das zugehörige Alignment, dessen Ende durch die Position des maximalen Eintrags in der Matrix bestimmt ist, in Zeit $O(mn)$ berechnet werden.
Bisher wurde immer auf die Berechnung des optimalen Alignments abgehoben. Jedoch sind im molekularbiologischen Kontext die Ähnlichkeitsmaße nicht derart,

daß durch Optimierung der daraus abgeleiteten Kostenfunktionen die aus Sicht des Anwenders biologisch relevante Lösung berechnet wird. Durch entsprechende Modifizierung der Ausleseprozedur erlaubt das obige Berechnungsschema jedoch nicht nur die Berechnung aller optimalen Alignments, sondern auch die Auswahl von suboptimalen Alignments [226, 247, 248], d.h. von Alignments, deren *Score* um ϵ von dem *Score* eines optimalen Alignments abweicht. Der Verwendung von lokalen und suboptimalen lokalen Alignments wird in späteren Kapiteln eine zentrale Rolle zukommen. Suboptimale Alignments werden in der RDP–Methode nicht nur verwendet, um Probleme zu reduzieren, die durch die Ungenauigkeit biologischer Bewertungssysteme entstehen, sondern auch, um den gefundenen Teillösungen eine Signifikanz beziehungsweise Zuverlässigkeit [227, 351] zuzuordnen.

Für spezielle Anwendungen hat die dynamische Programmierungsmethode vielfältige Erweiterungen erfahren, zum Beispielzum Vergleich von Sequenzen mit regulären Ausdrücken [244], und zur Einführung von Nebenbedingungen [56].

Während der allgemeine Ansatz der dynamischen Programmierungsmethode Insertionen und Deletionen erlaubt, suchen andere Methoden wie `BLAST` (**B**asic **L**ocal **A**lignment **S**earch **T**ool) [10] und `FASTA` [207, 274] nach zwei zusammenhängenden Sequenzabschnitten maximaler Ähnlichkeit. Beide Methoden basieren auf dem Vergleich von Sequenzfragmenten konstanter Länge (k–Tupel), unterscheiden sich aber in der Art und Weise wie die ähnlichen oder gleichen (k–Tupel) gefunden und zu Sequenzabschnitten größerer Länge kombiniert werden.

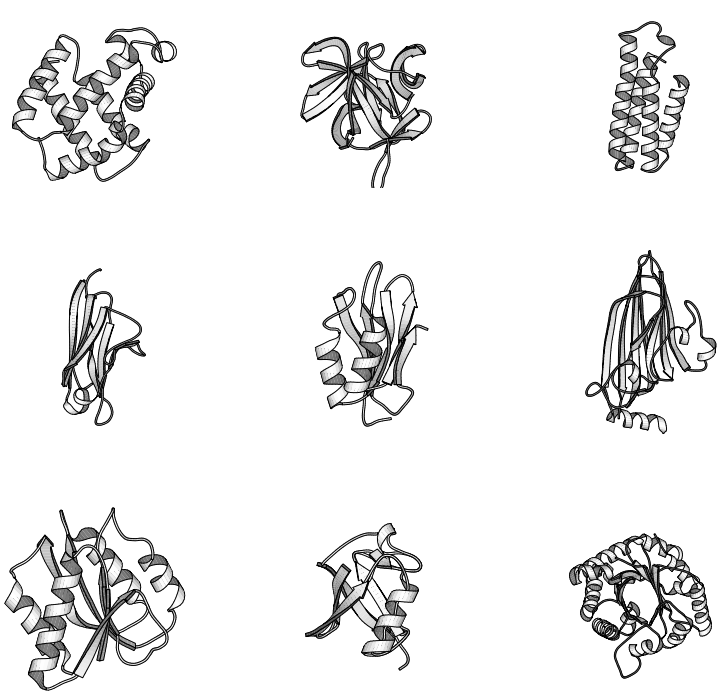
In `BLAST` werden in einer Vorabprozessierung der Datenbank alle oder nur die signifikanten (k–Tupel) entweder in einer *Hash*–Tabelle abgelegt oder in einem endlichen deterministischen Automaten codiert, der genau jene (k–Tupel) akzeptiert. Im Folgeschritt wird versucht, die so gefundenen Paare in beide Richtungen auszudehnen, bis der Wert des erweiterten Bereiches den bisher erreichten optimalen Wert um ein bestimmtes Maß unterschreitet.

In `FASTA` werden zunächst die Verschiebungen gleicher (k–Tupel) zwischen den zu vergleichenden Sequenzen bestimmt, dann werden heuristisch nah beieinander liegende (k–Tupel) gleichen Offsets unter Einbeziehung dazwischen liegender Positionen zu Sequenzabschnitten zusammengefaßt. Die endgültige Liste der gefunden Ähnlichkeiten ergibt sich nach Bewertung mit einer der `PAM`–Matrizen [73] (siehe Abschnitt 5.1).

Der Einsatzschwerpunkt dieser Methoden liegt eigentlich dort, wo eine Sequenz oder eine Menge von Sequenzen effizient gegen einen großen Datenbestand zu vergleichen sind. In der in dieser Arbeit entwickelten RDP–Methode können die mit diesen Methoden gefundenen maximal ähnlichen Sequenzfragmente aber auch als eine Möglichkeit zur Auswahl des Startpunktes des rekursiven Abstiegs verwendet werden (siehe Kapitel 6).

4.2 Faltungserkennung

Gegeben: • Bibliothek von Faltungen $\mathcal{F}$ • Sequenz



```
MASYPGHQHASAFDQAARSRGH
SNRRTALRPRRQQEATEVRPEQ
KMPTLLRVYIDGPHGMGKTTTT
QLLVALGSRDDIVYVPEPMTYW
RVLGASETIANIYTTQHRLDQG
EISAGDAAVVMTSAQITIGMPY
AVTDAVLAPHIGGEAGSSHAPP
PALTLIFDRHPIAALLCYPAAR
YLMGSMTPQ....
```

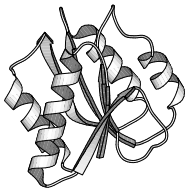
Gesucht: Plausible Faltung  + Sequenzstruktur –alignment

Abbildung 4.5: Proteinfaltungserkennung: Skizze der Problemstellung.

Die Ansätze der vergleichenden Modellierung (siehe Abschnitt 4.1) basieren auf der Erkennung von Verwandtschaften zu bereits aufgeklärten Proteinstrukturen mit den Methoden des Sequenzalignments. Diese Methoden können nur dann zuverlässige Ergebnisse liefern, wenn signifikante Sequenzähnlichkeiten vorhanden sind und sich diese zusätzlich über große Bereiche des Alignments erstrecken. Wie jedoch Strukturvergleiche und Klassifizierungen wie `SCOP` und `CATH` (siehe Abschnitt 3.3.5) belegen, sind auch jenseits signifikanter Sequenzidentitäten Ähnlichkeiten oder sogar evolutionäre Beziehungen zwischen Proteinen zu entdecken. Die Aufdeckung derartiger Beziehungen von bislang nicht strukturell aufgeklärten Proteinen zu bekannten Proteinstrukturen beziehungsweise Proteinfaltungen ist

das Ziel des im folgenden beschriebenen Ansatzes der Faltungserkennung.
Abbildung 4.5 veranschaulicht die bei der Faltungserkennung zu bearbeitende Problemstellung. Etwas formaler kann das Faltungserkennungsproblem, wie folgt definiert werden:

Definition 4.2 (Faltungserkennung)

Gegeben:
- *eine Sequenz* $A =< a_i >$ *mit* $a_i \in \Sigma$ *und* $A \in \mathcal{S}$, *wobei* $\mathcal{S}$ *die Menge der Proteinsequenzen ist*
- *eine Bibliothek von Faltungen* $\mathcal{F} \subseteq \{Proteinstrukturen\}$
- *ein Abstandsmaß* $\Phi : \mathcal{S} \times \mathcal{F} \longrightarrow I\!R$, *über das die Plausibilität einer Faltung für eine bestimmte Sequenz quantifiziert*

Gesucht: *Eine Faltung* $F \in \mathcal{F}$ *für die gilt:*

$$\Phi(A, F) = \min_{G \in \mathcal{F}} \quad \Phi(A, G)$$

In der vergleichenden Modellierung (siehe Abschnitt 4.1) wird die negative Sequenzidentität beziehungsweise Sequenzhomologie im Sinne der Definition 4.2 als Abstandsmaß Φ verwendet. Schon hier wird deutlich, daß die Bewertung eines Paares (A, F) stark von der Abbildung zwischen Positionen der Sequenz und Positionen der Faltung – hier berechnet durch ein Sequenzalignment (siehe Abschnitt 4.1.2) – abhängt. Im folgenden werden die Begriffe Abbildung und Alignment synonym gebraucht.
In dem Fall, daß die Beziehung zwischen der Sequenz und einer Faltung nicht mehr durch den Vergleich der Sequenzen ermittelt werden kann – also im eigentlichen Anwendungfall der Faltungserkennungsmethoden, entscheidet die Definition des Abstandsmaßes Φ sehr häufig über den Erfolg einer Methode. Naturgemäß unterscheiden sich die bisher in der Literatur vorgeschlagenen Verfahren neben der Art und Weise, wie die Abbildung zwischen Sequenz und bekannter Faltung erfolgt, hauptsächlich in der Definition von Φ. Bevor diese Unterschiede im Detail diskutiert werden, sollen im folgenden einige allgemeine Vorgehensweisen und Definitionen erörtert werden.
Die Ausgabe von Algorithmen zur Lösung des Faltungserkennungsproblems besteht in der Regel aus einer Rangliste $\mathcal{R}$, in der für eine untersuchte Sequenz A alle Faltungen F_i einer Bibliothek oder repräsentativen Menge $\mathcal{F}$ nach dem Abstandsmaß $\Phi(A, F_i)$ sortiert sind. Die zugrundeliegende Idee ist, daß je kleiner der Wert $\Phi(A, F_i)$ ist, desto wahrscheinlicher ist es, daß das Protein A die Faltung F_i annimmt.

Definition 4.3 (Rangliste)
Sei $A \in \mathcal{S}$, $F_i \in \mathcal{F}$ *und sei* Φ *ein Abstandsmaß im Sinne von Definition 4.2. Dann*

definiert Φ eine Ordnungsrelation $\leq_A$ auf der Menge der Faltungen $\mathcal{F}$ bezüglich einer Proteinsequenz A, mit

$$F_i \leq_A F_j \iff \Phi(A, F_i) \leq \Phi(A, F_j).$$

Die Rangliste $\mathcal{R} =< F_1, \ldots, F_n >$ *ist die gemäß dieser Ordnungsrelation $\leq_A$ sortierte Liste der $F_i \in \mathcal{F}$.*

Handelt es sich bei der Menge der Faltungen $\mathcal{F}$ um eine repräsentative Menge, so kann die Zuverlässigkeit einer Sequenz–Faltung–Zuordnung (A, F) anhand des Abstands des zugehörigen Abstandswertes $\Phi(A, F)$ von dem Mittelwert von Φ über alle Faltungen der Menge $\mathcal{F}$ bewertet werden. Der sogenannte *zscore* [280] ist das in der Regel hierfür benutzte statistische Abstandsmaß (siehe Definition 4.4).

Definition 4.4 (zscore)
Sei $\mathcal{R} =< F_1, \ldots, F_n >$ eine Rangliste. Der Mittelwert $\bar{\mathcal{R}}$ der Verteilung der Abstandswerte der Rangliste $\mathcal{R}$ ist dann definiert als

$$\bar{\mathcal{R}} = \frac{\sum_{i=1}^{n} \Phi(A, F_i)}{n}$$

und die Standardabweichung $\sigma_{\mathcal{R}}$ als

$$\sigma_{\mathcal{R}} = \sqrt{\frac{\sum_{i=1}^{n} (\Phi(A, F_i) - \bar{\mathcal{R}})^2}{n-1}}.$$

Der zscore *einer Zuordnung (A, F) ist dann definiert als*

$$zscore(A, F) = \frac{\bar{\mathcal{R}} - \Phi(A, F)}{\sigma_{\mathcal{R}}}$$

Der *zscore* gibt also den Abstand einer Zuordnung (A, F) vom Mittelwert der Verteilung der Abstandswerte in Einheiten der Standardabweichung der Verteilung $\sigma_{\mathcal{R}}$ an.
Wie bereits erwähnt, unterscheiden sich die aus der Literatur bekannten Lösungsansätze unter anderem

- durch die Definition und Modellierung der Faltung eines Proteins,
- durch die Definition des Abstandsmaßes zwischen einer Sequenz und einer Faltung beziehungsweise deren Modellierung und
- durch das algorithmische Verfahren zur Bestimmung der Abbildung zwischen Sequenz– und Strukturpositionen.

Naturgemäß sind diese Kriterien nicht unabhängig voneinander und bauen in vielen Fällen aufeinander auf, insbesondere haben die ersten beiden Punkte starken Einfluß auf die Anwendbarkeit verschiedener Lösungs– beziehungsweise Optimierungsmethoden. Abstrahiert man von der detaillierten Problemmodellierung, gibt es im wesentlichen zwei verschiedene Zielfunktionsmodelle:

Einkörperterme: In diesen Termen hängt die Bewertung einer Alignmentposition nur von der abgebildeten Aminosäure und lokalen Eigenschaften der Position in der Faltung ab. Beispiele für diese Art von Zielfunktionstermen sind die Aminosäureaustauschmatrizen (siehe Abschnitt 5.1) oder Umgebungspräferenzprofile (siehe Abschnitt 5.2.3).

Zwei– oder Mehrkörperterme: Bei Zwei– oder Mehrkörpertermen hängt die Bewertung einer Alignmentposition nicht nur von der auf diese Position abgebildeten Aminosäure, sondern auch von anderen Alignmentpositionen ab. Dies ist zum Beispiel der Fall für aminosäuretypabhängige Paarinteraktionspotentiale (siehe Abschnitt 5.2.4).

Gemeinsam ist allen Lösungsansätzen, daß sie die Plausibilität einer Faltung für eine gegebene Sequenz anhand einer Abbildung der Sequenz in das jeweils verwendete Strukturmodell, dem sogenannten Sequenzstrukturalignment beziehungsweise *Threading* bewerten. Der nächste Abschnitt gibt eine Definition des Sequenzstrukturalignmentproblems, das durch Abbildung 4.6 illustriert wird. Die Abschnitte 4.2.2 und 4.2.3 beschreiben zusammenfassend die in der Literatur vorgestellten Verfahren zur Faltungserkennung mit Ein– beziehungsweise Mehrkörpertermen.

4.2.1 Sequenzstrukturalignment

Als Sequenzstrukturalignment bezeichnet man die Abbildung einer Sequenz A in eine Struktur B. Da nicht alle Teile der Sequenz ein Äquivalent in der Struktur haben müssen und umgekehrt, wird das Sequenzstrukturalignment als injektiver, partieller Homomorphismus definiert, der die durch die Kettenverknüpfung der Proteinkette vorgegebenen Ordnungsrelationen sowohl auf Strukturseite ρ_B und auf Sequenzseite ρ_A erhält. Abbildung 4.6 zeigt einen derartigen Homomorphismus. Eine formale Problembeschreibung für das Sequenzstrukturalignment wird in Definition 4.5 gegeben.

Definition 4.5 (Sequenzstrukturalignment)
Sei $\mathcal{S}$ die Menge der Proteinsequenzen, $\mathcal{F}$ die Menge der Proteinfaltungen und $\mathcal{H}$ die Menge der Homomorphismen von $\mathcal{S}$ nach $\mathcal{F}$, dann ist das Sequenzstrukturalignment wie folgt definiert:

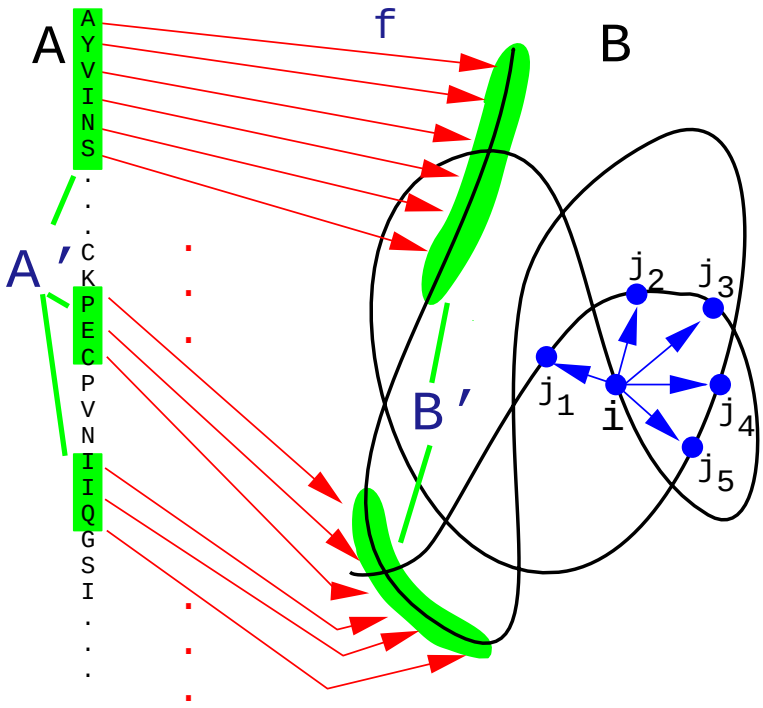


Abbildung 4.6: Sequenzstrukturalignmentproblem: gesucht ist ein partieller Homomorphismus f der Sequenz A in die Struktur B. Position i in Struktur B wechselwirkt mit den Positionen $j_1, \ldots, j_5$.

Gegeben:
- *eine Sequenz* $A =< a_i >$ *mit Ordnungsrelation* ρ_A *und* $a_i \in \Sigma$ *und* $A \in \mathcal{S}$
- *eine Struktur bzw. Faltung* $B \in \mathcal{F}$ *mit Ordnungsrelation* ρ_B
- *eine Kostenfunktion* $\phi : \mathcal{H} \times \mathcal{S} \times \mathcal{F} \longrightarrow I\!R$ *zur Bewertung der Abbildung* $f \in \mathcal{H}$ *einer Sequenz* S *in die Faltung* F

Gesucht: *partieller injektiver Homomorphismus* $f : A \longrightarrow B$ *mit*

$$\phi(f, A, B) = \min_{g \in \mathcal{H}} \quad \phi(g, A, B)$$

Um zu verdeutlichen, daß es sich bei ϕ in Definition 4.5 um eine Kostenfunktion zur Optimierung der Abbildung f handelt, wird hier das kleine ϕ anstelle des großen Φ verwendet (vergleiche Definition 4.2).

4.2.2 Faltungserkennung mit Einkörpertermen

Im weitesten Sinne handelt es sich auch bei Aminosäureaustauschmatrizen (siehe Abschnitt 5.1), die im Sequenzalignment (siehe Abschnitt 4.1.2) verwendet wer-

den, um Einkörperterme, da die Bewertung eines Austausches nur anhand der gerade betrachteten Position erfolgt. In diesem Zusammenhang sei daran erinnert, daß Faltungserkennung auch mit Sequenzalignmentmethoden, insbesondere mit eher strukturell motivierten Austauschmatrizen [164, 249] (siehe auch Abschnitt 5.1) möglich ist.
Werden Einkörperterme zur Faltungserkennung verwendet, so können die Sequenzstrukturalignments mittels sogenannter *Strukturprofilmethoden* berechnet werden, die häufig auch kurz als *Profilmethoden* bezeichnet werden. Profilmethoden sind aus der Sequenzanalyse bekannt, wo sie verwendet werden, um die Ähnlichkeit einer untersuchten Sequenz zu einer Gruppe von bereits alinierten Sequenzen zu bewerten [127, 128]. Dazu wird ein bereits berechnetes Alignment in eine positionsspezifische Kostentabelle – das *Profil* – umgerechnet, gegen die die neue Sequenz aliniert wird.
Ein *Strukturprofil* beschreibt dagegen eine Struktur durch eine Matrix der Länge der Struktur mit einer konstanten Anzahl von Attributen für jede Strukturposition. Unter dem Begriff *Profilmethoden* werden im Sequenzstrukturalignmentkontext diejenigen Methoden zusammengefaßt, die eine Sequenz gegen ein solches Strukturprofil alinieren.
Ein Eintrag in dem Strukturprofil kodiert dabei die Eigenschaften der zugehörigen Strukturposition, die zur Auswertung des verwendeten Einkörperterms (siehe Abschnitt 5.2.3) benötigt werden. Die Umgebungspräferenz gibt an, wie gut oder schlecht eine bestimmte Aminosäure in eine bestimmte Strukturumgebung paßt. Da die Bewertung einer Alignmentposition nur von der Aminosäure aus der Sequenz und einer Spalte des die Struktur beschreibenden Profils abhängt, können zur Lösung des Alignmentproblems die aus dem Sequenzalignment bekannten Algorithmen (siehe Abschnitt 4.1.2) eingesetzt werden. Die einzige Änderung besteht darin, daß der doppelt indizierte Zugriff in die Aminosäureaustauschmatrix (in Abschnitt 4.1.2 mit dem Funktionsnamen *sim* bezeichnet) in der innersten Schleife der dynamischen Programmierung durch eine unwesentlich komplexere Funktion ersetzt werden muß. In dieser Funktion sind die verschiedenen Einkörperterme für die Aminosäure der Sequenz anhand der die Strukturposition beschreibenden Attribute auszuwerten und gegebenenfalls unter Verwendung von Gewichtungsfaktoren aufzuaddieren.
Da die Beschreibung einer Strukturposition eine konstante Länge hat, wird die Zeitkomplexität der Profilmethoden wie im Sequenzalignment durch die verwendete Gapkostenfunktion bestimmt und ist daher maximal $O(n^3)$, in den typischerweise verwendeten Algorithmen für affine Gapkosten jedoch $O(n^2)$.
Die Existenz effizienter und zudem recht einfacher Algorithmen ist sicher ein Grund für die weite Verbreitung von Profilmethoden in der Faltungserkennung. Eine der ersten Profilmethoden wurde bereits 1991 von Bowie, Lüthy und Eisenberg vorgeschlagen [39]. In der Folgezeit sind in der Literatur weitere Profilmethoden vorgestellt worden [93, 209, 210, 221, 266, 267, 356, 357, 371, 374], die sich im wesentlichen in der verwendeten Strukturbeschreibung und den verwen-

deten Umgebungspräferenzen unterscheiden. Ein ausführlichere Beschreibung der verwendeten Einkörperterme gibt Abschnitt 5.2.3.
Die an der GMD entwickelte `123D`–Methode [5] ist eine Profilmethode, die auf den sogenannten Kontaktkapazitätspotentialen (`CCP`, siehe Abschnitt 5.2.3.2) basiert. Da die Laufzeit von `123D` $O(mn)$ (m Länge der Sequenz, n Länge des Strukturprofils) ist, können mit dieser Methode effizient auch größere Datenmengen, wie zum Beispiel die gesamte Proteinstrukturdatenbank, nach möglichen Faltungsvorlagen durchsucht werden. Daher wird diese Methode in den im Kapitel 8 beschriebenen Blindvorhersagen im ersten Schritt eingesetzt, um den Suchraum auf eine bestimmte Faltungsklasse einzuschränken (siehe Abschnitt 8.6).
Die Kontaktkapazitätspotentiale der `123D`–Methode werden aber auch beim Sequenzstrukturalignment mit der RDP–Methode als Bestandteil der Kostenfunktion (siehe Kapitel 7) verwendet, weil die Kontaktkapazitätspotentiale ein Maß für die Hydrophobizität darstellen und Hydrophobizität in empirischen Paarpotentialen allein durch hydrophobe Wechselwirkungen und damit nur unzureichend kodiert wird, da zum Beispiel Wechselwirkungen hydrophober Aminosäuren mit dem Lösungsmittel nicht modelliert werden.
Da es sich bei der `123D`–Methode um eine Profilmethode handelt, sind in `123D` auch alle in Abschnitt 4.1.2 beschriebenen Alignmentmodi möglich. Über den Alignmentmodus *lokal* ist es somit möglich, in einer untersuchten Sequenz nach Abschnitten zu suchen, die gut in eine Kontaktkapazitätsumgebung einer gegebenen Struktur passen. Die Suche nach biochemisch sinnvoll einander zuzuordnenden lokalen Bereichen in Sequenz und Struktur ist einer der elementaren Schritte der RDP–Methode. Die `123D`–Methode stellt als *Orakel* (siehe Abschnitt 6.2) in RDP eine der Möglichkeiten dar, diese Bereiche aufzufinden.
Eine wesentliche Weiterentwicklung der Profilmethoden besteht in der Einbeziehung von Sekundärstrukturvorhersagen (siehe Abschnitt 4.3.1), die mittlerweile – insbesondere wenn ein multiples Alignment der untersuchten Sequenz mit verwandter Sequenzen vorliegt – ein relativ hohes Maß an Zuverlässigkeit erlangt haben. Während Russell *et al.* [299] die Sekundärstrukturvorhersage [288] noch als einen zusätzlichen Filter nutzen, integrieren Rost *et al.* [292] und Rice und Eisenberg [284] die Vorhersage [288] der Sekundärstruktur und der Lösungsmittelzugänglichkeit in ihre jeweilige Profilmethode [209, 266]. Diese Erweiterung erfolgt methodisch dadurch, daß nicht nur die Struktur sondern auch die Sequenz durch ein Profil beschrieben wird. Das Sequenzprofil enthält neben der Aminosäuresequenz zusätzlich für jede Position die vorhergesagte Sekundärstruktur und Lösungsmittelzugänglichkeit, die bei der Auswertung des verwendeten Einkörperpotentials berücksichtigt werden.

4.2.3 Faltungserkennung mit Mehrkörpertermen

Bei der Faltungserkennung werden auch Zwei– und Mehrkörperterme zur Bewertung von Sequenzstrukturpaaren eingesetzt. Im Unterschied zu Einkörperter-

men bewerten Zwei– und Mehrkörperterme Wechselwirkungen zwischen zwei oder mehr Aminosäureresten, die in der Struktur „benachbart“ in der Sequenz aber beliebig weit von einander entfernt sein können (vergleiche Abbildung 4.6). Die Bewertung einer Wechselwirkung hängt dabei nicht nur von ihrer Art, sondern auch von dem Aminosäuretyp der diese Wechselwirkung ausbildenden Aminosäurereste ab. Daher werden diese Terme auch als *aminosäuretypabhängige* Wechselwirkungspotentiale bezeichnet.
Eine Alignmentbewertung mit derartigen Potentialen setzt – wie auch bei Einkörpertermen – die Kenntnis der Abbildung der Sequenz auf die Struktur voraus. Ist diese bekannt, so kann die Alignmentbewertung in Zeit linear in der Anzahl der zu bewertenden Wechselwirkungen erfolgen.
Lathrop [193] hat bewiesen, daß das Sequenzstrukturalignmentproblem *NP–vollständig* ist, wenn aminosäuretypabhängige Wechselwirkungspotentiale Bestandteil der Kostenfunktion sind und Insertionen beziehungsweise Deletionen beliebiger Länge erlaubt sind. Der NP–Vollständigkeitsbeweis erfolgt durch Reduktion des NP–vollständigen `ONE-IN-THREE-3SAT` [108] auf das dem Sequenzstrukturalignmentproblem zugeordnete Entscheidungsproblem, ob es ein Sequenzstrukturalignment mit einer Bewertung kleiner als ein vorgegebener Wert gibt.
Die NP–Vollständigkeit bedeutet, daß die Berechnung des optimalen Sequenzstrukturalignment nach heutiger Vermutung nur durch Algorithmen erfolgen kann, die im schlechtesten Fall eine Laufzeit exponentiell in der Länge der Problembeschreibung haben. In dieser Situation gibt es prinzipiell zwei Lösungsansätze:

Abstraktion der Problemschreibung: Durch die Wahl der Problembeschreibung wird das eigentlich zu lösenden Problem soweit vereinfacht, daß für das resultierende Problem in vertretbarer Laufzeit (zum Beispiel mit *Branch & Bound*–Verfahren [195]) eine optimale Lösung gefunden werden kann. Ein Beispiel für diese Vorgehensweise sind die *fragmentbasierten Sequenzstrukturalignmentverfahren* oder auch *fragment threading* genannten Verfahren, die im Anschluß genauer beschrieben werden.

Heuristiken: Die Problembeschreibung wird nicht vereinfacht, aber die Optimalitätsbedingung für die bestimmte Lösung des Sequenzstrukturalignmentproblems wird abgeschwächt. Hier sind in den vergangenen Jahren einige Lösungen vorgeschlagen worden, die im Anschluß ebenfalls kurz vorgestellt werden.

Die in dieser Arbeit vorgestellte RDP–Methode gehört zu den heuristischen Verfahren. Die Entscheidung, auf die Garantie der optimalen Lösung zu verzichten, wurde bewußt getroffen, da alle in der Proteinstrukturvorhersage verwendeten Bewertungssysteme nur eine grobe Approximation der die Faltung eines Proteins bestimmenden Kräfte sind. Eine biologisch sinnvolle Lösung des Sequenzstrukturalignmentproblems sollte daher nicht auf die absolute Optimierung einer sta-

tischen Kostenfunktion abheben, sondern adaptiv auf die im jeweiligen Kontext biologisch relevanten und damit signifikanten Signale eingehen.
Im folgenden werden beide Strategien und daraus resultierender Lösungsansätze sowie ihre Vor– und Nachteile diskutiert.

4.2.3.1 Fragmentbasierte Sequenzstrukturalignmentverfahren

Das fragmentbasierte Sequenzstrukturalignment wurde 1993 von Bryant und Lawrence eingeführt [48]. Fragmentbasierte Sequenzstrukturalignmentverfahren beschreiben eine Struktur nicht durch alle Aminosäurereste der Struktur, sondern reduzieren darüber hinaus die Proteinstruktur zu einem Modell, das aus sogenannten Kernfragmenten oder auch Kernfaltungsmotiven [48] besteht. Das bedeutet, daß nicht alle durch das verwendete Zwei– oder Mehrkörperpotential bewertbaren Wechselwirkungen bei der Berechnung des Sequenzstrukturalignments berücksichtigt werden, sondern nur die zwischen Aminosäureresten, die sich in unterschiedlichen Kernfragmenten befinden.
Die in diesen Methoden verwendete Faltungsbibliotheken bestehen in der Regel aus den Kernfaltungsmotiven, die als gemeinsame Faltungsmotive in homologen Proteinstrukturen identifiziert wurden (siehe Abbildung 4.7) und so als Repräsentanten der bekannten Faltungsfamilien dienen. Da diese Kernfaltungsmotive nicht nur eine, sondern mehrere homologe Strukturen repräsentieren, werden sie auch als *common cores* bezeichnet.
Beim fragmentbasierten Sequenzstrukturalignment werden die Kernfragmente in der Regel als die Sekundärstrukturelemente definiert [48, 194, 195]. Für die Einschränkung auf Sekundärstrukturelemente spricht, daß sie zwischen homologen Strukturen strukturell stärker konserviert sind als die Schleifenbereiche, wenn man allein auf die Anzahl der in einer Superposition einander zuzuordnenden Positionen achtet. Alle Schleifenbereiche, die in der Regel die aktiven Stellen der Proteine bilden, werden in diesem Modell nicht repräsentiert.
Bei der Berechnung der Sequenzstrukturalignments werden Gaps nur am Ende von Kernfragmenten erlaubt. Schleifenbereiche werden nur durch untere und obere Schranken für die Anzahl der Aminosäurereste in die Modellierung des Problems einbezogen, die zwischen zwei Kernfragmente für deren Verknüpfung übrig bleiben müssen, das heißt nicht mit Kernfragmenten aliniert werden dürfen.
Durch diese zwei Modellannahmen reduziert sich das Sequenzstrukturalignmentproblem, wie in Abbildung 4.7 gezeigt, darauf, die jeweils erste Position eines Kernfragmentes auf eine Position in der Struktur abzubilden. Da keine Gaps innerhalb von Kernfragmenten erlaubt sind, wird mit der Abbildung der ersten Position gleichzeitig auch die Abbildung der weiteren Positionen des Sekundärstrukturelements festgelegt. In der Literatur werden im wesentlichen drei Methoden zur Lösung des so beschriebenen Sequenzstrukturalignmentproblems vorgeschlagen:

- Vollständige Aufzählung [48].

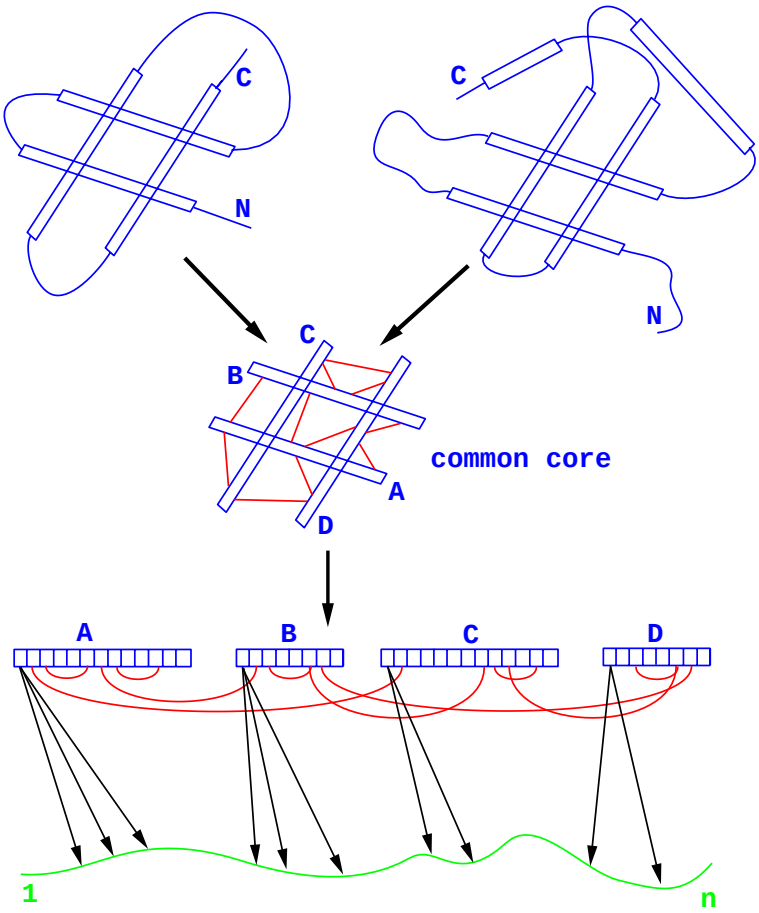


Abbildung 4.7: Fragmentbasiertes Sequenzstrukturalignment (*fragment threading*): homologe Proteinstrukturen werden in der Faltungsbibliothek gemeinsam durch einen *common core* beschrieben. Das Sequenzstrukturalignment bildet die Anfänge der Faltungsmotive auf Sequenzpositionen ab.

- *Monte Carlo* beziehungsweise *Gibbs Sampling* [46].
- *Branch & Bound* [194, 195].

Durch die Beschränkung auf Kernfragmente und das Verbot von Gaps in Kernfragmenten reduziert sich die Anzahl der möglichen Alignmentalternativen so stark [47], daß Bryant und Lawrence [48] die optimale Lösung des Sequenzstrukturalignmentproblems durch vollständige Aufzählung bestimmen können. Als Bewertungsfunktion verwenden Bryant und Lawrence dabei das in Abschnitt 5.2.4 beschriebene Paarpotential ohne zusätzliche Bestrafungsterme für Gaps. Wenn man Wechselgaps, die durch die Beschränkung des Alignments auf Kernfragmente naturgemäß entstehen, außer acht läßt, sind die berechneten Sequenzstrukturalignments dennoch kompakt, das heißt die Anzahl der zusätzlichen Insertionen

und Deletionen ist begrenzt, da große Gaps durch die oberen und unteren Schranken für die Anzahl der Resten in Schleifenbereichen ausgeschlossen sind.
Die vollständige Aufzählung ist jedoch nur dann möglich, wenn das Ziel die Berechnung einiger weniger Sequenzstrukturalignments ist, und nicht – wie bei der Faltungserkennung – eine Vielzahl von Alignments berechnet werden muß. Daher schlägt Bryant [46] eine heuristische Methode vor, die auf *Monte Carlo* beziehungsweise *Gibbs Sampling* beruht. Die Laufzeit für dieses Verfahren gibt er mit ungefähr 16 Minuten pro Alignment auf einer Silicon Graphics mit R4400 Prozessor an [46]. Auch er sieht die Probleme, die durch die Beschränkung der Problembeschreibung auf Kernfragmente entstehen [46], und versucht diesen zu begegnen, indem er die Erweiterung der Kernfragmente zur Laufzeit der Sequenzstrukturalignmentberechnung erlaubt. Dadurch steigt nicht nur die Verwendbarkeit der Alignments für eine anschließende vergleichende Modellierung, sondern zum Beispiel bei Globinen auch der *Score* und damit die Wahrscheinlichkeit der Erkennung eines korrekten Sequenzstrukturpaares [46].
Lathrop und Smith [194, 195] stellen ein effizientes *Branch & Bound*–Verfahren zur Lösung des fragmentbasierten Sequenzstrukturalignmentproblems vor. Wenn es in den durch die Rechnerresourcen gesetzten Grenzen terminiert, garantiert dieses Verfahren, die optimale Lösung für ein Sequenzstrukturalignmentproblem mit Mehrkörperpotentialen in der Kostenfunktion zu finden. Bei einem Test der Methode auf den paarweisen Sequenzstrukturvergleichen einer repräsentativen Menge von 58 Proteinen (eine Domäne, Monomere) mußte die Berechnung eines Sequenzstrukturalignments auf einem `DEC alpha 3000-M8000`–Rechner in 4% der Fälle nach zwei Stunden abgebrochen werden. In 0.5% der Fälle überschritt der Speicherplatzbedarf die Rechnerresourcen [195].
Das Verfahren basiert wesentlich auf den speziellen Eigenschaften der fragmentbasierten Beschreibung der Proteinstruktur, wie sie von Bryant und Lawrence [48] beschrieben wird (siehe oben):

- Die Größe des exponentiell wachsenden Suchraumes wird durch die Anzahl der Kernfragmente bestimmt [195], und nicht durch die wesentlich größere Anzahl der Aminosäurereste.

- Da alle Kernfragmente abgebildet werden müssen und Mindestlängen für die verbindenden Schleifen verwendet werden, können sowohl untere als auch obere Schranken für die in zulässigen Alignments erlaubten Anfangspositionen von Kernfragmenten in der Sequenz angegeben werden:

  - Bei der Alignmentberechnung kann der Suchraum auf die in diesem Sinne zulässigen Alternativen eingeschränkt werden.
  - In Faltungserkennungsexperimenten werden Sequenzstrukturpaare direkt ausgeschlossen, für die diese Bedingungen nicht erfüllbar sind. (Eine Erkennung von Teilstrukturen ist damit nicht mehr möglich.)

- Daß in Kernfragmenten keine Gaps erlaubt sind, wird wie folgt ausgenutzt:
  - Ein Alignment kann Speicherplatz sparend durch einen Zahlenvektor mit der Länge der Anzahl der Kernfragmente ($\#_{KF}$) beziehungsweise Sekundärstrukturelemente vollständig beschrieben werden.
  - Eine Menge von Alignments wird durch $\#_{KF}$ viele Intervalle repräsentiert, die die in dieser Menge erlaubten frühesten und spätesten Anfangspositionen von Kernfragmenten in der Sequenz festlegen. Auf diesen Intervallen basiert die im *Branch*–Schritt durchgeführte Aufteilung der zu einem Zeitpunkt der *Branch & Bound*–Prozedur noch zu untersuchenden, erlaubten Alignments.
  - Aussagekräftige untere Schranken für Mengen von Alignments werden effizient über die Summe der Minima der erlaubten Konfigurationen je zwei im Adjazenzgraph benachbarter Kernfragmente berechnet. Die effiziente Berechenbarkeit und die Güte der Schranken sind notwendige Voraussetzung für das Ausschließen von Teilen des Lösungsraums, ohne die zugehörigen Lösungen explizit berechnet zu haben (*Bounding*).
  - Die Berechnung noch schärferer unterer Schranken, die auch die Zulässigkeit der zur Berechnung der Schranken verwendete Alignmentkonfigurationen einbezieht [195], ist ohne die Einschränkungen des fragmentbasierten Modells nicht möglich.

Daraus wird deutlich, daß es sich bei dem von Lathrop und Smith vorgestellten Verfahren [194, 195] um eine speziell auf die Begebenheiten des fragmentbasierten Sequenzstrukturalignments zugeschnittene Lösungsstrategie handelt, die beim Wegfall der durch die gewählte Beschreibung des Sequenzstrukturalignmentproblems bedingten Einschränkungen nicht anwendbar ist. Beim nichtfragmentbasierten Sequenzstrukturalignment sind die für die Anwendbarkeit der Methode zu treffenden Annahmen nicht oder nur schwer möglich.
Das *Branch & Bound*–Verfahren garantiert die Optimalität der gefundenen Lösung bezüglich des verwendeten Bewertungssystems. Dies ist aber keinesfalls gleichbedeutend mit der Aussage, daß die Methode das bestmögliche Alignment einer Sequenz zu einer Struktur findet, wobei die Güte eines Alignments – vorausgesetzt beide Strukturen sind bekannt – zum Beispiel an der *RMS*–Abweichung der Superposition gemäß des Alignments gemessen wird.
Die Ursache für diese Diskrepanz zwischen Optimalität und Güte einer gefundenen Lösung ist zum einem in dem dem fragmentbasierten Sequenzstrukturalignment zugrundeliegenden Modell, zum anderen aber auch in der Ungenauigkeit der verwendeten statistisch abgeleiteten Bewertungssysteme zu suchen. Diese Bewertungssysteme stellen eine derart grobe Approximation dar, daß die an biologischen und strukturellen Kriterien gemessene bestmögliche Lösung nur selten mit der optimalen Lösung übereinstimmt (siehe dazu auch [131]).

Im der Originalarbeit [195] scheitert die *Branch & Bound*–Methode daher bereits bei der Reproduktion der Alignments zwischen der Sequenz und der zugehörigen nativen Struktur. Bei diesem einfachen Test werden nur 48% der Helices und 34% β–Stränge ohne Verschiebungsfehler (siehe Abschnitt 8.3.1) aliniert, für 18 beziehungsweise 24% ist der Absolutbetrag der Verschiebung der Sekundärstrukturelemente gegenüber der nativen Zuordnung größer als 9 Aminosäurereste [195]. Nach Aussage von Lathrop und Smith zeigen die fünf von ihnen getesteten Bewertungsfunktionen [48, 212, 232, 321, 361] nahezu identische Ergebnisse.
Bei der Anwendung der Methode auf biologisch relevante Anwendungsbeispiele bestätigen sich diese Probleme. Eine korrekte Zuordnung der Helices zweier homologer Globine (Hämoglobin `1dxt` Kette `A` und Myoglobin `1mbn`) liefert die Methode nur dann, wenn zuvor eine kurze Helix, die nur in einem der Proteine vorhanden ist, von Hand aus dem Strukturmodell entfernt wurde. Dieses Problem offenbart deutlich eine der Schwächen des zugrundeliegenden fragmentbasierten Sequenzstrukturalignmentmodells. Diese Schwäche resultiert daraus, daß in der Regel gefordert wird, daß allen vor der Berechnung des Alignments definierten Fragmenten ein Sequenzbereich zugeordnet wird, und so die Methode sehr sensibel auf bei der Fragmentdefinition gemachte Fehler reagiert.
Das Problem der Optimierung der falschen Kostenfunktion wird noch deutlicher bei Anwendung der Methode auf ein Paar von Proteinen, deren Ähnlichkeit nur strukturell und nicht mehr evolutionär interpretierbar ist [195]. In einer ersten Analyse [217] des `CASP II`–Wettbewerbs (siehe auch Abschnitt 8.5) bestätigen sich diese Ergebnisse. Gemäß dieser Analyse erkennen Lathrop und Smith zwar noch in 2 von 7 Fällen die korrekte Faltung, aber in keinem der Fälle gelingt es, mit der Methode ein Sequenzstrukturalignment zu berechnen, das die in dieser Analyse sehr schwach definierten Kriterien für ein korrektes Alignment erfüllt.

4.2.3.2 Allgemeine Sequenzstrukturalignmentverfahren

In nichtfragmentbasierten Verfahren wird die Proteinstruktur nicht nur durch Wechselwirkungen zwischen Aminosäureresten in Sekundärstrukturfragmenten beschrieben, sondern durch alle gemäß der Definition des verwendeten Mehrkörperbewertungssystems in der Struktur auftretenden Wechselwirkungen. Dabei werden Schleifenbereiche nicht durch die Definition des Problems aus der Problembeschreibung eliminiert, sondern gleichwertig mit Sekundärstrukturbereichen behandelt. Wie bereits erwähnt, werden die aktiven Stellen eines Proteins in aller Regel durch Aminosäuren in Schleifenbereichen des Proteins ausgebildet. Abbildung 4.8 zeigt am Beispiel der Familie der Serinproteasen, daß alle drei die katalytische Triade ausbildenden Aminosäuren (Serin 195, Histidin 57 und Asparaginsäure 102) in Schleifenbereichen der Struktur angeordnet sind. Ein Modell, das zur Funktionsvorhersage verwendet wird, muß daher in jedem Fall zumindest die für die Funktion des Proteins wichtigen Schleifenbereiche enthalten.
Zusätzlich ist die Anordnung dieser Reste im Raum über alle Mitglieder der Fami-

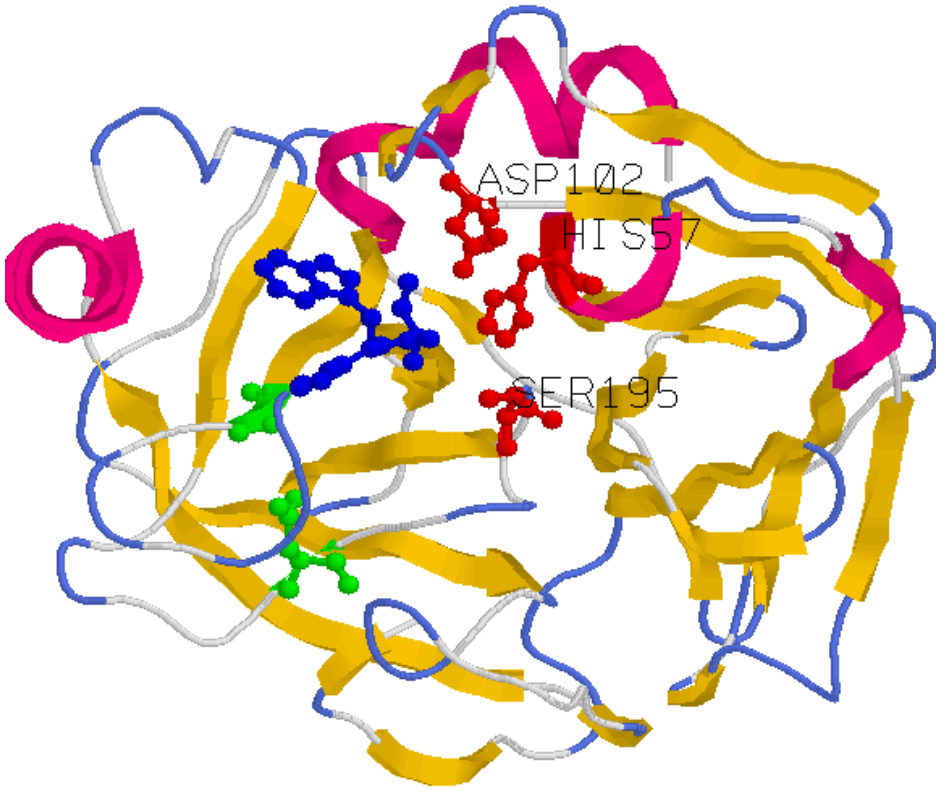


Abbildung 4.8: Bedeutung von Schleifenbereichen für die Funktion eines Proteins am Beispiel der Serinprotease: katalytische Triade (rot), Substratbindestelle (blau) und Spezifitätstasche (grün).

lie der Serinproteasen hoch konserviert. Und das gilt, obwohl die Sequenzidentität zwischen zahlreichen der Serinproteasen unter 20% liegt. Ähnliche Beobachtungen gelten für viele der Proteinfamilien, zum Beispiel auch für die Dehydrogenasen und fast alle ATP–bindenden Proteine (siehe dazu auch Abschnitt 8.6).

Daher können Schleifenbereiche wesentlich zur Erkennung von Ähnlichkeiten zwischen einer Sequenz und einer Struktur durch Verfahren beitragen, deren Kostenfunktionen auf der Tatsache basieren, daß die Struktur höher konserviert ist als die Sequenz. Aus diesem Grund wird in nichtfragmentbasierten Sequenzstrukturalignmentverfahren, wie in Abbildung 4.9 angedeutet, eine Struktur sowohl durch ihre Sekundärstrukturbereiche als auch Schleifenbereiche beschrieben und es werden nicht nur die wenigen Wechselwirkungen zwischen den Sekundärstrukturelementen, sondern auch die Wechselwirkungen zwischen Sekundärstrukturelementen und Schleifenbereichen und in Schleifenbereichen modelliert.

Ziel nichtfragmentbasierter Verfahren ist die Identifizierung möglichst aller Bereiche, für die eine Abbildung von Aminosäureresten der Sequenz auf Positionen der Struktur sinnvoll ist. Dazu müssen möglichst viele Aminosäurereste sowohl

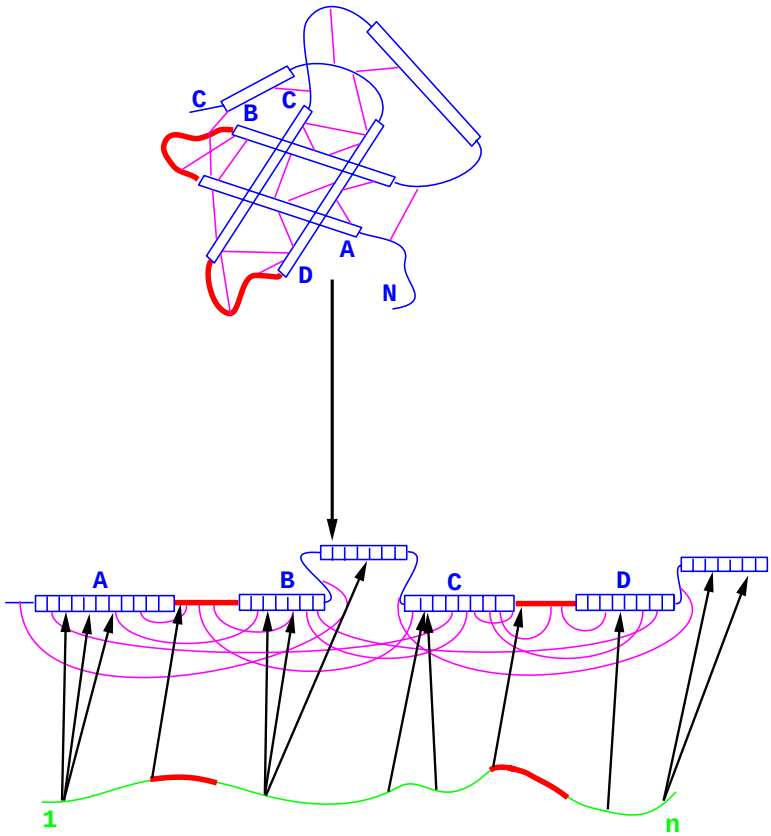


Abbildung 4.9: Allgemeines Sequenzstrukturalignment (*full threading*): Das Strukturmodell enthält zusätzlich die Schleifenbereiche und damit die Kontakte zu und zwischen Aminosäuren in diesen Bereichen. Das Sequenzstrukturalignment bildet möglichst große Teile der Sequenz auf dieses Strukturmodell ab.

der Struktur als auch der Sequenz berücksichtigt werden. Somit muß aber auch das Einfügen von Gaps an beliebigen Positionen und nicht nur an Enden von Kernfragmenten erlaubt sein. Außerdem sollten Schleifenbereiche in dem Alignment nicht nur als Wechselgaps modelliert werden. Damit entfallen für die Anwendbarkeit der für das fragmentbasierte Sequenzstrukturalignment vorgestellten Verfahren wesentliche Vereinfachungen des Problems:

- Eine Konfiguration im Lösungsraum des Sequenzstrukturalignmentproblems kann nicht mehr nur durch die Angabe der Abbildung des ersten Aminosäurerestes eines Kernfragmentes auf eine Sequenzposition beschrieben werden. Diese Vereinfachungen der Problembeschreibung ist eine der wesentlichen Voraussetzungen für die Anwendbarkeit der fragmentbasierten Sequenzstrukturalignmentverfahren.

- Außerdem können Längenbedingungen für Kernfragmente und Schleifenbe-

reiche nicht mehr zum Ausschließen nicht zulässiger Alignments und damit großer Bereiche des Lösungsraums verwendet werden.

Eine Verwendung der beschriebenen *Branch & Bound*–Verfahren zur Lösung des nichtfragmentbasierten Sequenzstrukturalignmentproblems würde implizieren, daß jedes Fragment aus einer einzelnen Strukturposition bestehen würde, da eine Zusammenfassung zu größeren zusammenhängenden Fragmenten in diesem Modell *a priori* nicht möglich ist. Damit entfällt die für die effiziente Anwendbarkeit dieser Verfahren notwendige Voraussetzung, das Problem auf die Plazierung längerer zusammenhängender Fragmente reduzieren zu können.
Beim nichtfragmentbasierten Sequenzstrukturalignment wird daher in der Regel auf die Garantie der Optimalität der gefundenen Lösung zugunsten einer geringeren Laufzeit verzichtet. Somit handelt es sich bei den in der Literatur vorgeschlagenen Verfahren durchweg um heuristische Verfahren, die den Sequenzalignment– beziehungsweise Strukturalignmentverfahren entlehnt sind. Folgende Verfahren werden vorgeschlagen:

- Abbildung der Sequenz auf die Struktur ohne Gaps [212, 325],
- Iterative Berechnung mit der Methode der dynamischen Programmierung unter Verwendung der sogenannten *eingefrorenen Approximation* [100, 118, 321, 365],
- Doppelte dynamische Programmierung [168, 167] und
- Rekursive dynamische Programmierung [340].

Die Abbildung der Sequenz auf die Struktur ohne Gaps ist für das Finden entfernt verwandter Strukturen sicher unzureichend, da nicht zu erwarten ist, daß die Anzahl der Aminosäurereste in allen Sekundärstruktur– und vor allem Schleifenbereichen zwischen entfernt homologen Proteinen erhalten bleibt. Die Methode der *rekursiven dynamischen Programmierung* (RDP) wird in dieser Arbeit vorgestellt. Daher sollen im folgenden nur die beiden anderen Ansätze kurz charakterisiert werden:
Die Kernidee der von Godzik *et al.* [118] mit dem Namen der *eingefrorenen Approximation* bezeichneten Vorgehensweise ist es, beim Alignment Kontakte nicht anhand der Kontaktpartner in dem aus dem Alignment resultierenden Modell, sondern anhand der in der Struktur gefundenen Kontaktpartner zu bewerten, also bei der Bewertung eines Kontaktes den jeweils zweiten oder dritten Kontaktpartner einzufrieren. Der Vorteil dieser Vorgehensweise ist, daß die Kontakte und ihre Bewertung durch Mehrkörperpotentiale, wie von den Einkörperpotentialen bereits bekannt, in einem Profil kodiert werden können. Damit kann die Berechnung des Alignments als das Alignment einer Sequenz gegen ein Profil effizient mit den verschiedenen Algorithmen erfolgen, die auf der Methode der dynamischen Programmierung basieren. Der nicht zu vernachlässigende Nachteil

besteht darin, daß nicht der Mehrkörperpotentialscore des aus dem Alignment resultierenden Modells optimiert wird. Vielmehr werden bei der Berechnung des Alignments Aminosäuren als Kontaktpartner verwendet, die bei typischen Anwendungsbeispielen, wo die Sequenzidentität unter 25% liegt, im resultierenden Modell nur selten erhalten sind.

Zur Behebung dieses Nachteils wird vorgeschlagen [118, 365], die Berechnung des Alignments mit einem angepaßten Strukturmodell zu wiederholen. Dazu wird die Sequenz gemäß des mit der eingefrorenen Approximation berechneten Alignments in die Struktur eingebettet und mit den so erhaltenen neuen Kontaktpartnern ein neues Profil erzeugt, gegen das dann die untersuchte Sequenz erneut aliniert wird. Dieser Prozeß wird solange iteriert, bis sich zum Beispiel der Score des Sequenzstrukturalignments nur noch unwesentlich ändert [365] beziehungsweise der Prozeß konvergiert.

Die *doppelte dynamische Programmierung* ist, wie der Name bereits andeutet, ein zweistufiges Verfahren. Auf beiden Stufen wird dabei das Verfahren der dynamischen Programmierung verwendet [168]. Dieses Verfahren wurde vom Strukturalignment [264, 339] adaptiert und in dem Programm `Threader` [168] implementiert. Bevor in der dynamischen Programmierung der obersten Stufe analog zum Sequenzalignment entschieden wird, ob Sequenzposition i auf Strukturposition j abgebildet wird, wird in der darunterliegenden Stufe ebenfalls mittels dynamischer Programmierung die optimale Belegung der Kontaktpartner von j durch die verbleibenden Aminosäurereste der Sequenz unter der Annahme bestimmt, daß Sequenzposition i auf Strukturposition j abgebildet wird. Damit wird bei der Entscheidung, i und j zu alinieren, immer von der zu diesem Zeitpunkt optimal möglichen Aminosäurebelegung der Kontaktpartner im späteren Modell ausgegangen. Dies ist eine Annahme, die durch das endgültige Alignment nicht erfüllt werden kann. Daher handelt es sich auch bei diesem Verfahren um eine Heuristik zur Lösung des Sequenzstrukturalignmentproblems. Die Laufzeit des Verfahrens beträgt bedingt durch die zweistufige dynamische Programmierung mindestens $O(n^4)$, wobei n das Maximum der Längen der Sequenz und der Struktur ist.

Ein Vorteil aller Verfahren, die auf der dynamischen Programmierung basieren, ist es, daß Kostenfunktionsbestandteile, die in den Profilmethoden verwendet werden, auch direkt bei der Optimierung von Sequenzstrukturalignments bezüglich von um Mehrkörperpotentialanteile erweiterten Kostenfunktionen einfließen können. So werden zum Beispiel in `Threader2` [167] neben Paarpotentialen auch Austauschmatrizen und Sekundärstrukturpräferenzen verwendet.

Ein wesentlicher Nachteil, der durch die RDP–Methode weitgehend behoben wird, besteht darin, daß bei der Alignmentberechnung mit Methoden der dynamischen Programmierung immer auch die Gapkosten eine entscheidende Rolle spielen. Die Festlegung der geeigneten Gapparameter ist dabei schwieriger als beim Sequenzalignment und damit ein bislang nicht zufriedenstellend gelöstes Problem.

4.3 *Ab initio*–Methoden

Wie bereits erwähnt, ist es Ziel der *ab initio*–Methoden, Proteinstrukturen oder –eigenschaften ohne die Kenntnis von Homologien zu bereits strukturaufgeklärten Proteinen vorherzusagen. Die *ab initio*–Methoden können wiederum in Abhängigkeit der vorherzusagenden Eigenschaften und der dazu eingesetzten Methoden in folgende Kategorien unterteilt werden:

Sekundärstrukturvorhersagemethoden versuchen die Elemente der Proteinsequenz den verschiedenen Sekundärstrukturelementen (α–Helix, β–Strang, Schleifenbereiche) zuzuordnen. Die Vorhersagen dieser Methoden beinhalten häufig auch Aussagen über die prognostizierte Lösungsmittelzugänglichkeit von Aminosäureresten.

Energieminimierungs– und Molekulardynamikmethoden versuchen die Proteinfaltung oder zumindest Teilaspekte davon auf der Basis empirischer Kraftfelder zu simulieren.

Kontaktvorhersagemethoden versuchen Kontakte in der Struktur aus der Sequenz vorherzusagen. Aszódi *et al.* [15] versuchen Kontakte aus Mustern konservierter hydrophober Aminosäurereste vorherzusagen, die auf der Grundlage multipler Sequenzalignments bestimmt werden. Diese Art evolutionärer Information wird auch genutzt, um Kontakte aus korrelierte Mutationen vorherzusagen [344, 116, 251, 259, 315]. Hubbard nutzt sie, um Kontakte zwischen β-Strängen vorherzusagen [156]. Über die Korrelation zwischen Sequenzmustern und strukturellen Motiven versucht Selbig nicht nur die Existenz von Kontakten, sondern auch deren Nichtexistenz vorherzusagen [312, 313].

Naturgemäß liegt der Einsatzschwerpunkt der in diesem Bereich entwickelten Methoden dort, wo die anderen Methoden zur Proteinstrukturvorhersage versagen, zum Beispiel, wenn die Datenbanken keine homologen Strukturen enthalten oder diese Homologien nicht mit den zur Verfügung stehenden Methoden bzw. Vergleichskriterien erkannt werden können. Aber sie bilden auch eine ideale Ergänzung zu den anderen im Bereich der Proteinstrukturvorhersage eingesetzten Methoden. So werden zum Beispiel Energieminimierungs– und Molekulardynamikmethoden bei der Nachoptimierung von mit den Methoden der vergleichenden Modellierung erzeugten Modellstrukturen verwendet [305], die Ergebnisse von Sekundärstrukturvorhersagen können sowohl zur Kontrolle der von Faltungserkennungsexperimenten als auch zu ihrer Steuerung eingesetzt werden.
Die Kontaktvorhersage befindet sich noch im Anfangsstadium. Daher wird im folgenden nicht näher daraufeingegangen. Aber falls ein Vorhersage von Kontakten oder auch von Nichtkontakten zuverlässig möglich ist, resultieren daraus Abstandsbedingungen, die von der in dieser Arbeit entwickelten RDP–Methode als zusätzliche Randbedingungen verwendet werden können.

4.3.1 Sekundärstrukturvorhersage

Bei der Sekundärstrukturvorhersage wird in der Regel ein Dreizustandsmodell (**H***elix*, **S***trand*,**C***oil*) zugrundegelegt, und zur Bewertung der Korrektheit einer Vorhersage wird daher das sogenannte Q3–Maß verwendet. Das Dreizustandsmodell hat unter anderem zur Folge, daß verschiedene Arten von Helices wie zum Beispiel 3_{10}–Helix und α–Helix, wie sie bei der Beschreibung aufgeklärter Strukturen unterschieden werden [173], unter dem Sammelbegriff Helix zusammengefaßt werden.
Im folgenden werden einige Sekundärstrukturvorhersagemethoden etwas genauer betrachtet. Dabei wird eine Klassifizierung der verschiedenen Ansätze nach der zugrundeliegenden Methodik versucht. Sofern nicht anders angegeben wird die Korrektheit der Vorhersage immer bezüglich der aus der Struktur abgeleiteten DSSP–Sekundärstrukturzuordnung [173] als Referenz angegeben. Die DSSP–Sekundärstrukturzuordnung ist zwar nicht in allen Fällen unumstritten [65], hat sich jedoch gegenüber anderen Verfahren [104, 206, 285, 326] durchgesetzt.

Statistische Ansätze: In diese Kategorie fällt einer der ersten vorgestellten Ansätze zur Sekundärstrukturvorhersage, der 1974 von Chou und Fasman erstmals vorgeschlagen wurde [60, 61, 94]. Statistische Ansätze basieren auf der statistischen Ableitung von Präferenzen einzelner Aminosäuretypen für bestimmte Sekundärstrukturelemente (in der Regel Helix, Strang, Schleife). Diese Präferenzen werden über die Sequenz aufgetragen und die eigentliche Zuordnung der Sekundärstrukturelemente erfolgt in der Originalarbeit von Chou und Fasman über eine Menge empirischer Regeln, die unter anderem die Sekundärstrukturpräferenzen der benachbarten Reste mit in Betracht ziehen. Als Schleifenbereich werden entweder die Bereiche definiert, die keiner Sekundärstruktur zugeordnet werden konnten, oder sie werden über einen eigenen Präferenzwert festgelegt [159, 336].

Fenstertechniken: Um die Abhängigkeit der Sekundärstrukturzuordnung des aktuell betrachteten Aminosäurerestes von seinen sequentiellen Nachbarn bei der Vorhersage einzubeziehen, werden bei den Fenstertechniken Sequenzfragmente fester Länge betrachtet, deren Zentrum der aktuelle Rest ist. Garnier und Robson [109, 110] bestimmen den Informationsbeitrag, den die Reste eines Fensters dazu beitragen, daß der zentrale Rest des Fensters in einer bestimmten Sekundärstrukturkonformation auftritt. Diese Informationsbeiträge werden aufsummiert und die Sekundärstruktur für den betrachteten Rest wird durch die zugehörige Majorität bestimmt. Die Korrektheit der verschiedenen Modifikation der sogenannten GOR–Methode wird von den Autoren mit $54 - 63\%$ angegeben [94].

Andere Ansätze verwenden die Verträglichkeit von Sequenzfragmenten mit aus der Datenbank entnommenen gleich langen Fragmenten von Umgebungsprofilen [301], wie sie zum Beispiel in einigen Faltungserkennungs-

verfahren [39] verwendet werden, oder mit für die verschiedenen Sekundärstrukturklassen abgeleiteten sogenannten *Sequenztemplates* [377], deren Positionen mit der Tendenz bestimmter Aminosäurereste für diese Position annotiert sind.

Periodizität: Bei der Vorhersage der Position von α–Helices kann die Eigenschaft ausgenutzt werden, daß die Reste einer typischen Helix in der Regel in einem bestimmten wiederkehrendem Muster lösungsmittelzugänglich und vergraben sind. Diesen Sachverhalt nutzen Blundell *et al.* [83] aus, indem sie zunächst aus dem multiplem Alignment homologer Sequenzen ein Profil gemäß umgebungsabhängiger Ersetzungstabellen ableiten und dieses Profil mit Hilfe der Fouriertransformation auf Periodizitäten hin untersuchen. Eine Helix wird vorhergesagt, wenn die gefundene Periodizität verträglich mit der für Helices typischen Periodizität ist. Mit dieser Methode kann außerdem die Orientierung der Helix, d.h. welche Seite der Helix dem Lösungsmittel zugewandt ist, vorhergesagt werden.

***Hidden Markov Models*(HMM):** Goldman *et al.* adaptieren den aus dem Bereich des Sequenzalignments bekannten HMM–Ansatz [86, 157, 278] für die Sekundärstrukturvorhersage [122], indem sie für jede Sekundärstrukturklasse ein gesondertes Markovprozeß–Modell verwenden.

Lineare Diskrimination: King und Sternberg [179] benutzen die statistische Methode der linearen Diskrimination, um aus den verschiedenen die Sekundärstruktur eines Restes bestimmenden Kriterien eine Vorhersage abzuleiten. Mögliche Kriterien sind zum Beispiel die der GOR–Methode entlehnten Sekundärstrukturtendenzen [109] oder die hydrophoben Momente des betrachteten Restes, die unter der Annahme berechnet werden, daß sich dieser Rest in einer Helix oder in einem Strang befindet. Nachdem die so gewonnene Vorhersage in einer Nachverarbeitung durch einfache bedingte Ersetzungsregeln überarbeitet wurde, wird eine Vorhersagegenauigkeit von im Schnitt 70.1% erreicht.

Multiple Alignments: Benner und Gerloff verwenden multiple Alignments homologer Sequenzen zur Sekundärstrukturvorhersage [27]. Anhand der Profile dieser multiplen Alignments analysieren sie die positionelle Divergenz der Sequenzen und leiten daraus unter Einbringung des „Verständnisses und der Intuition“ [27] der die Vorhersage durchführenden Person eine Sekundärstrukturvorhersage ab. Algorithmische Methoden dienen hier nur zur Unterstützung menschlichen Sachverstandes.

Neuronale Netzwerke: Die wohl bekannteste Methode zur Lösung des Sekundärstrukturvorhersageproblems basiert auf neuronalen Netzwerken. Diese 1993 von Rost und Sander vorgeschlagene Methode [288, 289] verwendet als Eingabe nicht einzelne Sequenzen sondern multiple Alignmentprofile und basiert auf der Hintereinanderschaltung von zwei *feed–forward*–Netzwerken,

die unabhängig mit standard Gradientenabstiegsmethoden [296] trainiert werden. Die Eingabeknoten des ersten Netzwerkes werden mit entsprechend der verwendeten Fensterbreite (zum Beispiel 13) vielen Häufigkeitsverteilungen der einzelnen Aminosäuretypen an der jeweiligen Position in dem betrachteten Fenster im multiplen Alignmentprofil als Eingabe versehen. Ausgabe dieses ersten Netzwerkes ist eine erste Sekundärstrukturzuweisung für den zentralen Rest des Fensters. Die Ausgabeknoten von einer beliebigen Fensterbreite (zum Beispiel 17) vielen Netzwerken der ersten Stufe bilden die Eingabe eines weiteren Netzwerkes, das sequenzunabhängig auf der Vorhersage der ersten Stufe arbeitet und zum Ziel hat, die Sekundärstrukturzuordnung von in der Sequenz zusammenhängenden Teilbereichen zu koordinieren. Um Fehler einzugrenzen, die durch die Wahl der Trainingsparameter entstehen können, arbeiten Rost und Sander mit einer Menge derartiger Netzwerkkaskaden und legen die endgültige Sekundärstrukturzuordnung per Mehrheitsentscheid der unabhängig trainierten Kaskaden fest. Mit dieser Methode wird eine Vorhersagekorrektheit von im Durchschnitt 71% [289] erreicht.

Eine Erweiterung ihres Ansatzes nutzen Rost *et al.* zur Vorhersage transmenbraner Helices [287]. Die Korrektheit ist in dieser Spezialanwendung mit 86% angegeben.

Ein weiterer neuronaler Netzwerkansatz wurde von Chandonia und Karplus [54, 55] vorgestellt. Sie verwenden standard *feed–forward*–Netzwerke, die aus zwei oder drei Lagen bestehen. Die Knoten zwischen je zwei Lagen sind vollständig miteinander verbunden. Als Aktivierungsfunktion verwenden sie eine typische Sigmoidalfunktion. Die Ausgabelage besteht aus je einem Knoten für eine Helix– und Strangvorhersage, die Eingabelage besteht aus 21 (Anzahl Aminosäuretypen plus leere Eingabe) mal Breite des um den vorherzusagenden Restes gewählten Fensters vielen Eingabeknoten. Durch Umstellung der Trainingsmethode von der *backpropagation*–Methode mit *steepest descent*–Optimierung [296] auf eine auf der konjugierten Gradientenminimierung basierende Methode [234] und der damit einhergehenden Möglichkeit des Trainings auf größeren Datenmengen konnte die Vorhersagekorrektheit von 62% [54] auf 67% [55] und durch zusätzliche Einbeziehung homologer Sequenzen [310] auf 72.9% gesteigert werden.

Selbstoptimierende Methoden Das von Geourjon und Deléage [112] vorgeschlagene `SOPM`–Verfahren bestimmt zunächst eine Menge von strukturaufgeklärten Proteinen, deren Sequenzen zu der untersuchenden Sequenz am ähnlichsten sind. Im zweiten Schritt werden dann alle 17 Reste langen Ausschnitte der Sequenz mit allen 17 Reste langen Peptiden der Proteinsequenzen dieser Menge verglichen. Falls der Vergleichswert einen bestimmten Grenzwert überschreitet, wird die im Vergleichsprotein gefundene Sekundärstruktur gewichtet mit dem Vergleichswert der zentralen Position

zugewiesen. Die Vorhersage ergibt sich dann aus dem maximalem Wert der aufsummierten Sekundärstrukturpräferenzen. Die Methode wird als selbstoptimierend bezeichnet, da die Gewichtungsparameter der oben beschriebenen Vorgehensweise zunächst iterativ kalibriert werden, indem sie so optimiert werden, daß auf der ausgewählten Teilmenge bekannter Strukturen die Vorhersage mit der wirklichen Sekundärstruktur möglichst gut übereinstimmt.

Die Einbeziehung homologer Sequenzen (`SOPMA`) [113] erhöht die Vorhersagekorrektheit auf einer 126–elementigen Testmenge [288] auf 69.5%. In Kombination mit der neuronalen Netzwerkmethode [291] kann für die 74% der von beiden Methoden identisch vorhergesagten Reste eine Genauigkeit von 82.2% erreicht werden.

Lokales Alignment: Salamov und Solovyev [302] nutzen lokale paarweise Sequenzalignments [359] zur Vorhersage von Sekundärstrukturelementen und erreichen dadurch auf einer typischen Testmenge [288] niederhomologer Proteine eine Genauigkeit von 71.2%. Dazu werden zunächst die 90 gemäß der Chou–Fasman Präferenzen für die Sekundärstrukturanteile [61] zur untersuchten Sequenz ähnlichsten strukturaufgeklärten Proteine ausgewählt. Für jedes Paar von untersuchter Sequenz und einem Protein aus der so bestimmten Menge werden etwa 50 lokale nicht überlappende lokale Alignments und deren Alignment*scores* berechnet [359]. Für jede Sequenzposition werden nun diese Alignments nach der Sekundärstruktur der ihr zugeordneten Strukturposition sortiert, die *Scores* der so entstandenen drei Teilmengen werden aufsummiert und die Sekundärstrukturvorhersage für die untersuchte Position wird durch die Teilemenge mit dem höchsten Summenscore festgelegt.

Kombinierte Verfahren: Frishman und Argos schlagen ein Verfahren vor [106], bei dem zunächst für die untersuchte Sequenz positionellen statistischen Tendenzen abgeleitet werden, die unter anderem die Tendenz zur Ausbildung von für bestimmte Sekundärstrukturen übliche Wasserstoffbrücken wiedergeben [105]. Im nächsten Schritt werden diese Tendenzen durch die statischen Tendenzen von Aminosäureresten angereichert, die in lokalen Alignments der jeweiligen Position zugeordnet werden, wobei eine zusätzliche Gewichtung durch ein Maß für das verwendete Alignment erfolgt. Auf der Testmenge [288] niederhomologer Proteine erreicht dieses Verfahren einen Korrektheitsgrad von 75%.

Multiple Alignments werden in Kombination mit allen bisher beschriebenen Verfahren benutzt, um die Vorhersage der Sekundärstruktur der betrachteten Sequenz auf eine breitere Wissensbasis zu stützen, indem zum Beispiel der *Score* für die Zugehörigkeit zu einem bestimmten Sekundärstrukturtyp durch Mittelwertbildung über alle zu den betrachteten Positionen alinierten Aminosäuren

berechnet wird [301]. Dabei sind Deletionen und Insertionen in multiplen Alignments ein sicherer Indikator für Schleifenregionen [78]. Durch die Einbeziehung evolutionärer Verwandtschaften wird der Korrektheitsgrad der Vorhersage in den meisten Fällen um $4-7\%$ gesteigert. Die algorithmische Einbeziehung multipler Alignments wird als die grundlegende Verbesserung der letzten Jahre angesehen [23]. Sie hat bewirkt, daß der Korrektheitsgrad auf in der Regel über 70% (in zuverlässigen Bereichen sogar auf bis zu 89%) gesteigert werden konnte. Damit hat die Sekundärstrukturvorhersage einen Genauigkeitsgrad erreicht, um als Startpunkt oder als Lieferant von Nebenbedingungen für die Tertiärstrukturvorhersage zu dienen.
Die oben beschriebenen Methoden stellen nur eine Auswahl der publizierten Ergebnisse dar, die sich auf die Beschreibung prinzipiell unterschiedlicher Vorgehensweisen beschränkt. Einige dieser Methoden werden nicht nur zur Sekundärstrukturvorhersage, sondern in gradliniger Erweiterung der Methodik auch zur Vorhersage der Lösungsmittelzugänglichkeit von Sequenzpositionen verwendet [290, 26].

4.3.2 *Ab initio*–Tertiärstrukturvorhersage

Trotz der in Abschnitt 3.1 beschriebenen Probleme bei der Auswertung der zu betrachtenden Energiefunktion wird versucht, Proteinstrukturen oder Eigenschaften von Proteinstrukturen mittels Energieminimierungs– und Molekulardynamikverfahren vorherzusagen.
Die verwendeten Kraftfelder – wie `GROMOS` [347], `Amber` [360] oder `CHARMM` [43] – stellen jedoch nur Näherungen dar, in denen entropische Anteile nur durch ein Wasserkontinuum mit hoher Dielektrizitätskonstante ϵ_0 einbezogenen werden. Die für die jeweilige Simulation verwendete Näherungsstufe hängt von der interessierenden Systemeigenschaft ab oder noch häufiger davon, für welche Näherung das betrachtete Problem mit den heutigen Verfahren und Rechnern noch lösbar ist. So ist zum Beispiel bei der Betrachtung chemischer Reaktionen eine quantenmechanische Behandlung der Freiheitsgrade der Elektronen unbedingt erforderlich. In diesem Falle steigt jedoch der Berechnungsaufwand pro Energiewert mindestens mit der dritten Potenz der Anzahl der Elektronen. Bei der Simulation des Verhaltens von Proteinen werden daher in der Regel die quantenmechanischen Freiheitsgrade aus der Systembeschreibung eliminiert. Das umgebende Lösungsmittel wird häufig nur implizit durch Anpassung der Wechselwirkungsfunktion berücksichtigt. Häufig wird sogar das Modell durch Zusammenfassung ganzer Atomgruppen (zum Beispiel Aminosäurereste) zu einer Kugel vereinfacht. Daher spricht man auch von einer Kraftfeldhierarchie [348]. Zur Laufzeitreduktion ist es vielfach notwendig, die Betrachtung der genannten Wechselwirkungen auf Atome in bestimmten Abstandsbereichen einzuschränken. Dies hat jedoch Diskontinuitäten in der Energiefunktion zur Folge, die zur ungewollten Erwärmung des Systems während der Simulation und damit zu fehlerhaften Ergebnissen führen

können. Die geeignete Zusammensetzung von Kraftfeldern wird in einer Reihe von Artikeln [25, 218] und Übersichten [44, 132, 345, 346, 348] diskutiert.
Neben dem verwendeten Kraftfeld unterscheiden sich die einzelnen Ansätze durch die Art und Weise, wie der Konformationsraum durchsucht wird. Dies geschieht mit

- systematischen Durchsuchungsmethoden, die den gesamten Konfigurationsraum des Moleküles durchsuchen, oder mit
- Methoden, die die Generierung eines möglichst repräsentativen Satzes von Konformationen zum Ziel haben. Diese wiederum unterteilen sich in
  - Methoden, die eine prinzipiell unkorrellierte Folge zufälliger Konformationen erzeugen, und
  - Methoden, die schrittweise die neue Konformation aus der oder den vorangegangenen Konformationen generieren. Hierzu gehören *Monte–Carlo*–, Moleküldynamikmethoden und Methoden der stochastischen Dynamik.

Eine Charakterisierung der verschiedenen Kraftfelder und Simulationsprotokolle gibt van Gunsteren in [348]. Eine ausführlichere Diskussion findet sich in [345, 346]. Allgemein läßt sich sagen, daß Simulationsmethoden heute nicht zur *ab initio*–Vorhersage von Proteinstrukturen geeignet sind, da ein System, in dem die Faltung eines Proteins angemessen simuliert werden könnte, aufgrund seiner Größe und der damit verbundenen Größe des Konformationsraums mit den heute zur Verfügung stehenden Methoden nicht durchsucht und bearbeitet werden kann. Mit genetischen Algorithmen ist es zum gegenwärtigen Zeitpunkt in einigen Fällen möglich, für 14 Aminosäuren lange Fragmente Konformationen zu generieren, die der nativen Konformation ähnlich sind [276].
Eine Sonderrolle nehmen die Gittermethoden ein, bei denen das Modell sowohl hinsichtlich der Beschreibung des Proteins als auch hinsichtlich des verwendeten Kraftfeldes soweit vereinfacht wird, daß Faltungsuntersuchungen im Rechner möglich werden. Komplexitätstheoretisch gesehen, ist das Proteinfaltungsproblem *NP*–vollständig unabhängig von dem zugrundeliegendem Gittermodell [28, 135]. Insbesondere bei der theoretischen Untersuchung der die Proteinfaltung bestimmenden Faktoren nehmen gitterbasierte Methoden in der Literatur großen Raum ein [21, 49, 177, 181, 231, 307, 308]. Unabhängig davon, ob kartesische Gitter [53, 69, 97, 196] oder tetraedrische Gitter [146] oder hierarchisch aufgebaute Gitter [186, 187] verwendet werden, wird die Proteinkette auf einen in das Gitter eingebetteten, kreuzungsfreien Pfad abgebildet [282, 300], bei dem jeder Punkt des Gitters in Abhängigkeit vom verwendeten Modell durch maximal ein Atom oder eine Aminosäure belegt wird. Ein sehr einfaches Faltungsmodell, an dem zahlreiche theoretische Untersuchungen zur Proteinfaltung durchgeführt wurden,

ist das *hydrophob–hydrophil*–Modell (HP–Modell) von Dill [80]. In diesem Modell gibt es nur zwei Typen von Aminosäuren, hydrophile und hydrophobe, und das Ziel ist, eine möglichst hydrophobe Packung in einem kartesischen Gitter zu erhalten. Für dieses Problem schlagen Hart und Istrail Approximationsalgorithmen vor, die in linearer beziehungsweise quadratischer Laufzeit eine Lösung garantieren, die höchstens um einen multiplikativen Faktor von $\frac{3}{8}$ von der optimalen Lösung entfernt liegt [136]. Sie erzielen ein ähnliches Resultat auch für ein vereinfachtes gitterfreies Modell [137]. Jedoch ist das zugrundeliegende HP–Modell so sehr vereinfacht, daß es für die Vorhersage von Proteinstrukturen in der Detailgüte von Röntgenstrukturen nicht verwendbar ist und die gefundenen Lösungen biologischen Gütekriterien nicht genügen.
Simulationsverfahren werden dort nutzbringend eingesetzt, wo es zum Beispiel um Bindungsmechanismen oder lokale Bewegungen des bereits gefalteten Proteins geht [176]. Beim Versuch der Erklärung von Wirkungsmechanismen darf nicht außer acht gelassen werden, daß sich ein Protein auch in nativer Konformation durch hohe Beweglichkeit insbesondere der dem Lösungsmittel exponierten Schleifen und Seitenketten auszeichnet. Ein gut untersuchtes Beispiel für die Notwendigkeit dieser Beweglichkeit ist die in Myoglobin eingebaute Hämgruppe, die der Speicherung von Sauerstoff dient. Wäre das Myoglobin starr in seiner kristallographisch aufgeklärten Konformation, wäre es dem Sauerstoff so gut wie unmöglich, die Hämgruppe zu erreichen oder wieder zu verlassen [176].
Ein weiterer Einsatzschwerpunkt in Bezug auf Proteine ist die Strukturverfeinerung von röntgenkristallographisch oder mit NMR–Methoden aufgeklärten Proteinstrukturen oder Proteinligandkomplexen [346]. Die experimentelle Strukturaufklärung kann systembedingt nur bis zu einer gewissen Genauigkeit erfolgen. Diese Ungenauigkeiten lassen jedoch auch energetisch ungünstige und damit sicher nicht reale Konformationen von zum Beispiel Seitenketten zu, die in einem nachfolgenden Verfeinerungschritt mittels Computersimulation eliminiert werden. Der *RMS*–Abstand der durch die Simulation erhaltenen Struktur zur Ausgangsstruktur liegt mit etwa 1Å in der gleichen Größenordnung, wie man sie erhält, wenn man unterschiedliche Kraftfelder verwendet [162].
Ausgehend von der nativen Faltung kann mittels molekular–dynamischer Simulation eine Vielfalt von möglichen Konformationen generiert werden, die eine Überdeckung des Konformationsraums vom gefalteten Protein bis zum ungefalteten Protein darstellen. Aus dieser Vielfalt leiten Boczko und Brooks [34] die Energieoberfläche der freien Faltungsenergie für ein kleines helikales Protein in Abhängigkeit von dem *Gyrationsradius* ab, der ein Maß für die Ausdehnung des Proteins ist. Aus dieser Simulation ergibt sich ein Unterschied von $11 kJmol^{-1}$ zwischen nativem und ungefaltetem Protein. Außerdem leiten die Autoren aus der Simulation die Existenz eines metastabilen Zustandes ab, der energetisch ungefähr $4.5 kJmol^{-1}$ und vom Gyrationsradius etwa um den Faktor 1.5 vom nativen Zustand entfernt ist. Derartige Simulationsexperimente tragen zum Verständnis des Faltungsprozesses bei, der experimentell nicht beobachtet werden kann.

Kapitel 5

Bewertungssysteme

Die im folgenden vorgestellten Bewertungssysteme bewerten Alignments als eine Summe von Einzelereignissen. Ein Einzelereignis kann dabei im Falle eines Sequenzalignments (siehe Abschnitt 4.1.2) darin bestehen, daß eine Aminosäure in dem Alignment konserviert ist oder durch eine andere ersetzt wird, oder im Falle des Sequenzstrukturalignments (siehe Abschnitt 4.2.1) darin, daß die durch das Alignment definierte Abbildung der Sequenz in die dreidimensionale Struktur bestimmte Aminosäuren in räumliche Nähe zueinander bringt. Die Bewertungssysteme ordnen diesen Einzelereignissen eine Bewertung zu, die eine Unterscheidung nach günstigeren und weniger günstigeren Ereignissen erlaubt.
Unabhängig davon, ob es sich um Bewertungssysteme für Sequenzalignments oder Sequenzstrukturalignments handelt, basieren fast alle vorgestellten Bewertungssysteme auf der Idee, anhand von Lern– beziehungsweise Trainingsmengen zu beobachten, wie oft ein bestimmtes Einzelereignis in dieser Menge vorkommt, und daraus Rückschlüsse auf die Wahrscheinlichkeit zu ziehen, dieses Ereignis in einem berechneten Alignment zu beobachten. Dabei wird immer davon ausgegangen, daß es sich um unabhängige Ereignisse handelt und sich so die Wahrscheinlichkeit eines Alignments als das Produkt der Einzelwahrscheinlichkeiten ergibt. Es ist klar, daß die Annahme, daß es sich um unabhängige Ereignisse handelt, in den meisten Fällen eine starke Vereinfachung der Realität darstellt. In der Regel werden die Wahrscheinlichkeiten logarithmiert und dann als *Score* oder *Pseudoenergie*, so daß sich der *Score* eines Alignments als Summe der Einzel*scores* berechnet.

5.1 Sequenzabhängige Bewertungssysteme

5.1.1 Aminosäureaustauschmatrizen

Die einfachste Art und Weise, ein rein sequenzabhängiges Bewertungssystem für Alignments zu definieren, ist es, Identitäten mit einem Wert $\gamma \in \mathbb{R}$ und Paare nicht identischer Aminosäuren mit einem Wert $\chi \in \mathbb{R}$ zu bewerten, wobei für ein sinnvolles Bewertungssystem $\gamma > \chi$ sein sollte und beide Werte in Abhängigkeit der Gapbestrafungsparameter α und β (vergleiche Definition 4.1) zu wählen sind. Für Nukleotidsequenzen stellt dieses einfache Modell eine geeignete Modellierung dar. Aufgrund der unterschiedlichen biochemischen Eigenschaften von Aminosäuren (siehe Kapitel 2) ist es bei Proteinsequenzen jedoch sinnvoller, Ersetzungen nicht uniform, sondern in Abhängigkeit von der Art der Ersetzung zu bewerten. Zu diesem Zweck versucht man, die Wahrscheinlichkeit $Prob(a \longrightarrow b)$

der Mutation einer Aminosäure $a \in \Sigma$ zu einer anderen Aminosäure $b \in \Sigma$ zu bestimmen. Eine direkte Ableitung der Mutationswahrscheinlichkeiten aus den biochemischen Eigenschaften ist weder möglich noch besonders sinnvoll, da das eigentliche Mutationereignis nicht auf Aminosäure– sondern auf Nukleotidebene stattfindet.
Ein weitverbreitetes Modell zur Abschätzung von Mutationswahrscheinlichkeiten sind die sogenannten PAM–Matrizen (**P***oint* **A***ccepted* **M***utations*), die erstmals von Dayhoff *et al.* [73] vorgeschlagen wurde. Für das PAM–Modell werden folgende Annahmen gemacht:

- Akzeptierte Mutationen sind Mutationen, die an existierenden Spezies zu beobachten sind. Das heißt, daß Mutationen, die von der Evolution nicht selektiert wurden, nicht berücksichtigt werden.
- Mutationen werden als direkte Mutationen und als ungerichtet angenommen. Das bedeutet, daß eine Mutation $a \longrightarrow b$ nicht über den Umweg einer dritten Aminosäure $c \in \Sigma$ entstanden ist, und daß $Prob(a \longrightarrow b)$ gleich $Prob(b \longrightarrow a)$ ist und somit die Information über die zeitliche Reihenfolge der Mutationen verlorengeht.

Die erste dieser Bedingungen ist dadurch erfüllt, daß die Lernmenge zur Ableitung der Mutationswahrscheinlichkeiten aus Alignments von existierenden Proteinsequenzen besteht. Die zweite Annahme ist weitaus schwieriger zu erfüllen. Einige Autoren versuchen ihr gerecht zu werden, indem sie nur Mutationen zwischen sehr eng verwandten Sequenzen zählen [73], andere [249] vernachlässigen die zweite Annahme, wiederum andere [142] verwenden bewußt auch Sequenzpaare mit niedrigerer Sequenzidentität.
Sei nun f_{ab} die Anzahl der beobachteten Mutationen $a \longrightarrow b$, sei p_a die Wahrscheinlichkeit einer Aminosäure a, dann ergibt sich die Mutationswahrscheinlichkeitsmatrix M [314]:

$$\begin{aligned} M_{aa} &= 1 - \frac{\sum_{a \neq b} f_{ab}}{\sum_a \sum_{a \neq b} f_{ab} \ * 100 \ p_a} \\ M_{ab} &= Prob(a \longrightarrow b) \\ &= \frac{f_{ab}}{\sum_a \sum_{a \neq b} f_{ab} \ * 100 \ p_a} \qquad \forall a \neq b \end{aligned}$$

Die Mutationswahrscheinlichkeitsmatrix M soll einen Evolutionsschritt modellieren, der willkürlich so definiert ist, daß in einem Evolutionsschritt im Durchschnitt eine von 100 Aminosäuren mutiert wird. Daher erfolgt die Normierung in der obigen Definition durch den Faktor 100 unabhängig von der Datenmenge, in der die Mutationsereignisse gezählt wurden. Die so definierte Matrix enthält also die Mutationswahrscheinlichkeiten für einen Evolutionsschritt. In der Regel ist jedoch anzunehmen, daß zwischen zwei verwandten Proteinen weitaus mehr

evolutionäre Schritte stattgefunden haben. Unter der Annahme, daß es sich bei den einzelnen Schritten um unabhängige *Markov*–Prozesse handelt, erhält man die entsprechende Matrix der Mutationswahrscheinlichkeiten durch die Bildung der k–ten Potenz der Matrix M, wobei k die Anzahl der angenommen Evolutionsschritte ist.
Um ein Bewertungsmaß $sim(a, b)$ für Alignments (vgl. Definition 4.1) zu erhalten, werden die Mutationswahrscheinlichkeiten M_{ab} auf die Wahrscheinlichkeit normiert, daß Aminosäure b zufällig an einer Alignmentposition auftritt, und abschließend logarithmiert. Damit ergibt sich die evolutionäre Ähnlichkeit zweier Aminosäuren a und b als:

$$sim^k(a,b) = \phi \ \log \frac{M^k_{ab}}{p_b} \qquad\qquad \phi \in I\!R$$

Betrachtet man den evolutionären Übergang von einer Sequenz in eine andere als Folge unabhängiger Mutationsereignisse, so ergibt sich die Wahrscheinlichkeit des Übergangs, dessen Resultat durch das Alignment der Sequenzen beschrieben wird, als Produkt der Einzelwahrscheinlichkeiten beziehungsweise als Summe ihrer Logarithmen.
Im folgenden wird die Matrix sim^{250} auch synonym als `PAM250` oder auch einfach *Dayhoff*–Matrix bezeichnet, wenn es sich um die von Dayhoff abgeleitete Matrix handelt.
Häufig wird aber auch die folgende, einfachere Definition der Mutationswahrscheinlichkeit M'_{ab} verwendet [249]:

$$M'_{ab} \ = \ \frac{f_{ab}}{f_a * f_b} \ ,$$

wobei mit f_a die bei der Zählung der Mutationen beobachtete Häufigkeit der Aminosäure a ist. Additive Ähnlichkeiten erhält man dann wiederum durch den Trick der Logarithmierung.
Der wesentliche Unterschied zwischen den in der Praxis eingesetzten Ähnlichkeitsmatrizen besteht in der Bestimmung der akzeptierten Punktmutationen:

- Dayhoff *et al.* [73] beobachten 1572 Mutationen in 71 Gruppen eng verwandter Proteine (mindestens 85% Sequenzidentität). Mutationen werden aus den multiplen Alignments und den zugehörigen phylogentischen Bäumen dieser eng verwandten Sequenzen abgelesen.
- Gonnet *et al.* [123] indizieren zunächst die Sequenzdatenbank in einem *Patricia tree* [236], der die effiziente Ermittlung verwandter Sequenzen ermöglicht. Im nächsten Schritt werden die so als verwandt identifizierte Sequenzen aliniert. Die aus den Alignments abgelesenen Mutationen verwenden sie als Basis für die Ableitung neuer Ähnlichkeitsmatrizen – auch bekannt als *Gonnet*–Matrizen.

- Einen ähnlichen Ansatz verfolgen Jones *et al.* [169]. Sie teilen zunächst die Datenbank in *Cluster* auf, deren Elemente eine Sequenzidentität größer als 85% haben, berechnen innerhalb der Teilmenge die Alignments, zählen die Mutationen und verwenden dann die gleichen statistischen Methoden wie Dayhoff [73]. Diese Matrizen werden im folgenden auch kurz als *Jones*–Matrizen bezeichnet.
- Während bei der Matrixableitung nach Dayhoff nur Mutationsereignisse zwischen eng verwandten Sequenzen gezählt und mehrere Evolutionsstufen über die Potenzierung der Ausgangsmatrix nachgebildet werden, legen Henikoff und Henikoff [142] verschiedene Schwellwerte prozentualer Sequenzidentitäten für die Auswahl von Basisalignments zugrunde. Außerdem verwenden Henikoff und Henikoff nur lokale Alignments ohne Gaps (sogenannte *blocks* [143]) als Datenbasis. Die Matrix `BLOSUM80` ist also zum Beispiel aus einer Datenbank mit Blöcken abgeleitet, deren Sequenzsegmente eine Sequenzidentität größer als 80% aufweisen. Für den Vergleich entfernter verwandter Proteine schlagen die Autoren die Verwendung von Matrizen mit niedrigerem Schwellwert vor. Vergleicht man Matrizen anhand ihres Informationsgehalts, so ist die `BLOSUM45` mit der `PAM250` und die `BLOSUM80` mit der `PAM120` vergleichbar [142].
- Naor *et al.* [249] berechnen Austauschmatrizen auf der Basis von Alignments, die sie aus der strukturellen Superposition strukturähnlicher Regionen nicht verwandter Proteine ableiten.
- Overington *et al.* erstellen eine zunächst auf der Basis einer Datenbank von Strukturalignments [306] umgebungsabhängige Austauschmatrizen [267], deren Integration zu einer allgemeinen Austauschmatrix [164] führt.

Eine grundlegende Annahme für die Anwendbarkeit der obigen Methodik ist, daß Mutationen immer unabhängig voneinander stattfinden. Dies ist in der Realität sicher nicht der Fall. Daher wird unter anderem versucht, die Idee der Austauschwahrscheinlichkeiten von einer Aminosäure auf k–Tupel zu übertragen [202]. Zum einen wächst dann aber die Größe der Matrizen exponentiell mit k, zum anderen nimmt die Anzahl der Zählungen mit steigendem k schnell ab, so daß statistische Methoden nicht mehr zuverlässig anwendbar sind. Ein möglicher Ausweg ist, statt Aminosäuren Klassen von Aminosäuren zu betrachten und den Detailgrad der Klasseneinteilung von der Entfernung zum zentralen Rest abhängig zu machen.
Alternativ wird auch versucht, Austauschmatrizen aus den strukturellen Eigenschaften einzelner Aminosäuren zum Beispiel über den Vergleich der Korrelation der ϕ/ψ–Winkelverteilungen einzelner Aminosäuren abzuleiten [253]. Diese Matrizen haben jedoch bis heute keine große Akzeptanz gefunden.
Besteht die Anwendung darin, eine neue Sequenz mit einem bereits existierendem multiplen Alignment zu vergleichen, ist die Erweiterung der Austauschmatrizen

auf positionsspezifische Bewertungsmatrizen sinnvoll [128, 141], die auch als Profile bezeichnet werden. Dabei wird die Häufigkeit einer Aminosäure an einer Position des bereits gegebenen multiplen Alignments zur Gewichtung der Einträge einer üblichen Austauschmatrix benutzt. Ein *Hidden Markov*–Modell [86, 189] kann in diesem Zusammenhang auch als eine positionsspezifische Bewertungsmatrix verstanden werden, die durch einen iterativen Alignmentprozeß entstanden ist, der zusätzlich die Gapbestrafungsterme während der Berechnung positionsabhängig festlegt.

5.1.2 Vergleich sequenzabhängiger Bewertungssysteme

Ein Vergleich der Güte der unterschiedlichen Austauschmatrizen ist schwierig, da beim Einsatz von Alignmentmethoden Insertionen und Deletionen möglich sind und damit die verwendeten Gapbestrafungsterme eine wichtige Rolle spielen.
Henikoff und Henikoff [144] umgehen das Problem der optimalen Wahl der Gapbestrafungsterme, indem sie die unterschiedlichen Matrizen anhand ihrer Sensitivität bei ihrer Verwendung in `BLAST` und `FASTA` vergleichen. Ihr Test besteht in der Erkennung von verwandten Proteinen, die den Listen funktionsverwandter Proteine der `PROSITE`–Datenbank [17, 20] entnommen sind. Sie kommen zu dem Ergebnis, daß sowohl die *Gonnet*–Matrix als auch die Matrizen der *Jones*–Serie besser sind als die Matrizen der *Dayhoff*–Serie. Die besten Ergebnisse jedoch erzielten sie mit der von ihnen selbst abgeleiteten `BLOSUM62`–Matrix. Vorteile für Matrizen [164, 267], die auf der Basis von Strukturalignments entstanden sind und mehr auf die biochemischen Eigenschaften der einzelnen Aminosäuren abheben, sehen sie in Bereichen niedriger Homologie, da diese üblicherweise nicht bei der Ableitung der anderen Matrizen berücksichtigt werden.
Die Untersuchungen von Pearson [273] bestätigen diese Ergebnisse. Darüberhinaus bezieht Pearson auch lokale Alignments [327] mit in den Vergleich ein. Dazu variiert er die Kosten für eine Insertion beziehungsweise Deletion in Einerschritten im Intervall $[-4, -1]$ und die Gaperöffnungskosten in Zweierschritten im Intervall $[-16, -6]$. Dennoch zeigen die Ergebnisse, daß der lokale Alignmentalgorithmus bei Verwendung der besten `BLOSUM`–Matrizen und der heuristisch angepaßten Gapkosten, sensitiver ist als die einfacheren `BLASTP`– und `FASTA`–Algorithmen.
Vogt *et al.* [354] bewerten die Qualität der verschiedenen Austauschmatrizen anhand der Fähigkeit von lokalen und globalen Alignmentalgorithmen, ein vorgegebenes strukturrichtiges Alignment [270] mit der jeweiligen Austauschmatrix reproduzieren zu können. Für jede Matrix werden dazu unabhängig die optimalen Gapkosten heuristisch mit einem hierarchischen Diskretisierungsansatz bestimmt. Untersuchungen zum parametrischen Alignment [352, 380] zeigen jedoch, daß die Wahl der geeigneten Gapkostenparameter grundlegend die Qualität von Alignments beeinflußt und eine Diskretisierung über ein Raster möglicher Werte nicht die Lösung des Problems darstellt. Dennoch zeigen sich in der Analyse von Vogt *et al.* die moderneren Matrizen, wie `BLOSUM` und *Gonnet*, auch unter dem Maß

der Alignmentqualität den älteren Matrizen überlegen. Diese Ergebnisse werden auch durch aktuelle Untersuchungen von Abagyan und Batalov [1] bestätigt.

5.2 Empirische Potentiale zur Faltungserkennung

Bei den Potentialen, die zur Faltungserkennung eingesetzt werden, handelt es sich in der Regel um statistisch abgeleitete Potentiale und nicht um physikalisch motivierte Potentiale, wie sie zum Beispiel in der Moleküldynamik eingesetzt werden (siehe Abschnitt 4.3.2), oder gar die freie Energie ΔG (siehe Abschnitt 3.2). In der statistischen Mechanik werden dies statistisch abgeleiteten Potential auch als *Potentials of mean force* [30] bezeichnet.

5.2.1 Ableitung empirischer Potentiale

Empirische Potentiale bewerten das Auftreten von Aminosäuren in bestimmten Konstellationen. Die verschiedenen, in der Literatur vorgeschlagenen empirischen Potentiale unterscheiden sich im wesentlichen darin, was unter dem Begriff Konstellation verstanden wird (siehe Abschnitte 5.2.3 und 5.2.4) und was in der folgenden Definition 5.1 zur Vereinfachung der Beschreibung der allgemeinen Herleitung empirischer Potentiale als Ereignis definiert wird.

Definition 5.1 (Ereignis)
Ein Ereignis e *ist das Auftreten eines Tupels* $[a] = (a_1, \ldots, a_k)$ *von* $k \in \mathbb{N}$ *Aminosäuren in einer bestimmten Konstellation. Die Menge aller Ereignisse wird als* Λ *bezeichnet.*

Bei der Faltungserkennung geben empirische Potentiale für ein Tupel $[a]$ von Aminosäuren die Präferenz dafür an, daß diese Aminosäuren zum Beispiel in der nativen Konformation eines Proteins in einer bestimmten Konstellation auftreten, beziehungsweise nach Definition 5.1 an einem Ereignis e beteiligt sind.
Potentiale für $k = 1$ werden als *Einkörperpotentiale* bezeichnet, für $k \geq 2$ spricht man von *Zwei–* beziehungsweise *Mehrkörperpotentialen*. Für die Faltungserkennung wird am häufigsten der Spezialfall $k = 2$ verwendet, dann spricht man auch von *Paarinteraktionspotentialen*. Die in dieser Arbeit verwendeten Einkörperpotentiale werden in Abschnitt 5.2.3, die Mehrkörperpotentiale in Abschnitt 5.2.4 genauer beschrieben.
Zur Ableitung dieser Potentiale werden zunächst die interessierenden Ereignisse $e \in \Lambda$ in einer repräsentativen Menge von Proteinen gezählt. Aus diesen Zählungen $g_{[a]}(e)$ für das Tupel $[a]$ werden im nächsten Schritt die relativen Häufigkeiten $f_{[a]}(e)$ für Ereignis e berechnet:

$$f_{[a]}(e) = \frac{g_{[a]}(e)}{\sum\limits_{e' \in \Lambda} g_{[a]}(e')}$$

Die relativen Häufigkeiten $f_{[a]}$ setzen die Häufigkeit, für das Tupel $[a]$ das Ereignis e zu beobachten, in Verhältnis zur Häufigkeit, ein beliebiges Ereignis aus der Menge Λ für $[a]$ zu beobachten. Sie werden als Approximation für die Wahrscheinlichkeitsverteilung der Ereignisse verwendet.
Um aus den relativen Häufigkeiten den Energiebeitrag des Tupels $[a]$ zur durchschnittlichen Energie für ein Ereignis e zu erhalten, werden sie in Verhältnis zu einem Referenzzustand $\hat{\pi}_{[a]}(e)$ gesetzt. Sippl [319] schlägt die Wahrscheinlichkeit, Ereignis e unabhängig von $[a]$ zu beobachten, als Referenzzustand vor, also:

$$\hat{\pi}_{[a]}(e) = \hat{f}(e) = \frac{\sum\limits_{[b]} g_{[b]}(e)}{\sum\limits_{[b]} \sum\limits_{e' \in \Lambda} g_{[b]}(e')}$$

Dieser Referenzzustand ist unabhängig von dem einzelnen Protein und definiert sich über die Aufsummierung über alle Proteine der Referenzmenge. Dagegen schlagen Bryant und Lawrence [48] einen proteinspezifischen Referenzzustand vor, der aus der durch Zählung der Ereignisse für Permutationen der jeweiligen nativen Sequenz berechnet wird, also:

$$\hat{\pi}_{[a]}(e) = \hat{f}'_{[a]}(e) = \frac{g'_{[a]}(e)}{\sum\limits_{e' \in \Lambda} g'_{[a]}(e')}$$

Wobei mit g' die aufsummierten Zählungen über die jeweiligen permutierten nativen Sequenzen bezeichnet sind. Für die in dieser Arbeit eingesetzten Paarinteraktionspotentiale konnte zumindest bei der Erkennung von nativen Sequenzstrukturpaaren, kein gravierender Unterschied zwischen der Verwendung verschiedener Referenzzustände festgestellt werden [366].
Nach dem *Boltzmanngesetz* ist die Wahrscheinlichkeitsdichteverteilung $p(e)$ eines Ereignisses e für ein physikalisches System im Gleichgewicht proportional zum Energiebeitrag $E(e)$ dieses Ereignisses:

$$p(e) = \frac{1}{Z} \quad \exp\left(-\frac{E(e)}{kT}\right)$$

Dabei ist $k = 1.381\ 10^{-23} JK^{-1}$ die *Boltzmannkonstante* und T die Temperatur. Z ist im Fall einer diskreten Zustandsmenge die Zustandssumme über alle Zustände aus Λ. Die Umstellung der Gleichung liefert das *inverse Boltzmanngesetz*:

$$E(e) = -kT\ ln\ p(e) - kT\ ln\ Z$$

Approximiert man nun die Wahrscheinlichkeitsdichteverteilung $p(e)$ durch die berechneten relativen Häufigkeiten $f_{[a]}(e)$ und setzt diese zudem in Beziehung zum gewählten Referenzzustand, so erhält man mit dem inversen Boltzmanngesetz

die relative Pseudoenergie $\Delta' E_{[a]}(e)$ für ein Tupel $[a]$ und ein Ereignis e nach folgender Gleichung:

$$\Delta' E_{[a]}(e) = -kTln\left(\frac{f_{[a]}(e)}{\hat{\pi}_{[a]}(e)}\right) - kTln\left(\frac{Z_{[a]}}{Z}\right)$$

Das Problem an dieser Gleichung ist, daß die Zustandssummen Z und $Z_{[a]}$ nicht aus den Wahrscheinlichkeitsdichteverteilungen berechenbar sind. Sippl argumentiert in [319], daß in erster Näherung $Z \approx Z_{[a]}$ und damit der zweite Teil der Differenz der obigen Gleichung ungefähr 0 ist.
Die Zustandssummen sind unabhängig von den bewerteten Ereignissen, und damit konstant für eine untersuchte Proteinsequenz sind. Bei der Faltungserkennung wird aber – sieht man von den nicht alinierten Teilen der Sequenz ab – die Sequenz festgehalten und in verschiedene Faltungen eingefädelt. Daher gilt in erster Näherung für alle Pseudoenergiebewertungen von Sequenzstrukturpaarungen der gleiche Offset und der konstante Term wird in der Regel bei der Berechnung der relativen Pseudoenergie $\Delta E_{[a]}(e)$ weggelassen:

$$\Delta E_{[a]}(e) = -kTln\left(\frac{f_{[a]}(e)}{\hat{\pi}_{[a]}(e)}\right)$$

Eine Struktur B kann als eine Menge Λ_B von Ereignissen e beziehungsweise von Paaren $([b], e)$ beschrieben werden, die aus einem Ereignis und einer Aminosäurebelegung $[b] = AS_B(e)$ der an diesem Ereignis e in B beteiligten Aminosäuren bestehen. Da die Pseudoenergieanteile unter der Annahme der Unabhängigkeit der Ereignisse aufgrund der Logarithmierung aufaddiert werden können, kann die Pseudoenergie einer Struktur B, wie folgt definiert werden:

$$\Delta E(B) = \sum_{([b],e):e\in\Lambda_B,[b]=AS_B(e)} \Delta E_{[b]}(e)$$

Ein Sequenzstrukturalignment $f : A \longrightarrow B$ (vergleiche Definition 4.5) bewirkt eine andere Instantiierung der Ereignisse aus Λ_B mit Aminosäuren. Dabei gibt es Ereignisse, für die durch f keine vollständige Belegung der an diesem Ereignis beteiligten Strukturpositionen vorgibt, das heißt $\exists\ b_i \in [b]\ f^{-1}(b_i) = \emptyset$. Zur Vereinfachung sei $f^{-1}([b]) = \emptyset \Longleftrightarrow \exists\ b_i \in [b]\ f^{-1}(b_i) = \emptyset$. In den meisten Anwendungen empirischer Potentiale wird $\Delta E_{\emptyset}(e)$ zu 0 gesetzt, was in der Terminologie des Sequenzalignments einem Verzicht auf Bestrafungsterme für Insertionen und Deletionen entspricht. Die Pseudoenergie $\Delta E(f, A, B)$ eines Sequenzstrukturalignment f der Sequenz A gegen die Struktur B ist dann definiert als:

$$\Delta E(f, A, B) = \sum_{([a],e)\ :\ e\in\Lambda_B\ \wedge\ [a]=f^{-1}(AS_B(e))} \Delta E_{[a]}(e)$$

Nachdem nun gezeigt wurde, wie man aus der Beobachtung definierter Ereignisse e in einer repräsentativen Menge von Proteinen ein aminosäuretypabhängiges Bewertungssystem zur Optimierung und Bewertung von Sequenzstrukturalignments

ableiten kann, werden in den folgenden verschiedene Klassen von möglichen Ereignissen definiert.
Dabei liegt der Schwerpunkt auf der Beschreibung der empirischen Potentiale, die in der RDP–Methode zur Optimierung von Sequenzstrukturalignment eingesetzt werden. Die in der Literatur vorgeschlagenen Pseudoenergiepotentiale unterscheiden sich untereinander und zu den an der GMD entwickelten Potentialen neben kleineren Unterschieden in der verwendeten Normierung im wesentlichen in der Art und Weise, wie eine Proteinstruktur oder Proteinfaltung in dem der Definition von zählbaren Ereignissen zugrundeliegenden Modell beschrieben wird. Der folgende Abschnitt befaßt sich mit dieser Art der Modellbildung.

5.2.2 Faltungsmodelle für empirische Potentiale

Zur Vereinfachung erfolgt die abstrakte Beschreibung einer Proteinstruktur beziehungsweise einer Proteinfaltung B im folgenden auf der Basis eines attributierten Graphen $G_B = (V_B, E_B, \mu, \nu)$. Dieser sogenannte Proteingraph beschreibt den topologischen Aufbau einer Proteinstruktur. Dazu werden sowohl kovalente Bindungen als auch attributierte Wechselwirkungen innerhalb des Proteins durch Graphkanten modelliert. Definition 5.2 gibt die klassische, physikalisch–motivierte Definition der Wechselwirkungsrelation in Proteinen.

Definition 5.2 (Wechselwirkungsrelation)
Aminosäurereste $[b] \in B$ stehen in Relation $\sim_x$ genau dann, wenn eine kovalente Bindung oder eine Kette kovalenter Bindungen in der Form von Peptidbindungen ($\sim_p$) oder Disulfidbrücken ($\sim_s$), oder wenn Wasserstoffbrücken ($\sim_h$), hydrophobe Wechselwirkungen ($\sim_{hp}$) oder elektrostatische Wechselwirkungen ($\sim_e$) zwischen den Aminosäurereste $[b]$ bestehen.

Während kovalente Bindungen einfach zu bestimmen sind, erfordert die Erkennung von Wasserstoffbrücken in einer Proteinstruktur wesentlich mehr Aufwand, da zunächst die in der Proteinstrukturen fehlenden Wasserstoffatome generiert und dann nach möglichen Donor–Akzeptor–Paaren gesucht werden muß. Die Definition von hydrophoben Wechselwirkungen kann in erster Näherung über geometrische Kriterien und die Auftrennung von Aminosäureresten in hydrophobe und polare Aminosäuren erfolgen. Aber spätestens bei der Analyse von elektrostatischen Wechselwirkungen muß eine Zuordnung von Teilladungen erfolgen.
Die Wechselwirkungen bedingen eine sehr genaue geometrische Anordnung der beteiligten Atome. Betrachtet man im Gegensatz dazu die *RMS*–Abweichungen schon zwischen verschiedenen Strukturen hoher Homologie (siehe Abbildung 3.4), kann man nicht davon ausgehen, daß die detaillierten Ausprägungen dieser Wechselwirkungen auch nach der Abbildung einer zu weniger als 25% homologen Sequenz A auf die Struktur B erhalten bleiben.
Die in der Literatur vorgeschlagenen empirischen Potentiale für die Faltungserkennung abstrahieren daher von dem Begriff der physikalischen Wechselwirkung,

indem sie eine Wechselwirkung über ein einfaches geometrisches Distanzkriterium definieren. Für die Faltungserkennung werden dabei seltener Potentiale auf atomarer Ebene [75, 323, 324] betrachtet. In der Regel heben die Faltungserkennungpotentiale auf den Abstand von Aminosäuren ab, wobei der euklidische Abstand zweier Aminosäuren zum Beispiel auf folgende Arten verwendet wird:

- als euklidischer Abstand der C_α–Atome [319],
- als Abstand der C_β–Atome [139],
- als Abstand sogenannter virtueller C_β–Atome [48], die durch Verschiebung des C_β–Atoms (zum Beispiel um 2.4Å) in Richtung des (C_α, C_β)–Vektors berechnet werden (das für Glycin fehlende C_β–Atom wird künstlich erzeugt),
- als der minimale euklidische Abstand zweier Nicht–Wasserstoff–Atome der Aminosäurereste,
- als der minimale euklidische Abstand der van–der–Waals–Kugeln zweier Nicht–Wasserstoff–Atome der Aminosäurerest.

Wenn im folgenden von dem Abstand $\|b_i, b_j\|$ zweier Aminosäuren b_i und b_j gesprochen wird, so ist der euklidische Abstand der C_β–Atome der Reste gemeint, sofern nicht anderes gesagt wird.

Definition 5.3 (Distanzkontaktrelation)
Sei $\|b_i, b_j\|$ der (euklidische) Abstand zweier Aminosäuren b_i, b_j in der Struktur B.
Dann stehen Aminosäurereste $[b] \in B$ für einen gegebenen Abstandsgrenzwert max_dist in Distanzkontaktrelation $\sim_d$ *genau dann, wenn*

$$\forall (b_i, b_j) \in [b] \ : \ \|b_i, b_j\| \ \leq \ max\_dist.$$

Der Abstand $max_{(b_i,b_j)\in[b]} \ \|b_i, b_j\|$ ist ein Attribut des Kontaktes $\sim_d [b]$.

Für die in der Literatur vorgeschlagenen empirischen Potentiale variiert der maximal berücksichtigte Abstandswert max_dist zwischen 10Å [48] und 20Å [319]. Bei der Betrachtung von paarweisen Distanzkontakten wird in der Regel ein Distanzkontaktrelation $\sim_d (b_i, b_j)$ mit dem Abstand $|i - j|$ der Aminosäuren in der Sequenz der Peptidkette attributiert.
Die Berechnung der Distanzkontaktrelationen für eine Struktur B benötigt im schlechtesten Fall eine Laufzeit $O(|B|^2)$, wobei $|B|$ die Anzahl der Aminosäuren von B ist. Im Mittel kann die Berechnungsdauer durch Einbeziehung geometrische Abstandsbedingung stark verkürzt werden. Zum Beispiel kann ausgenutzt werden, daß wenn zwei Aminosäurereste sehr weit voneinander entfernt sind, auch die unmittelbaren und mittelbaren Nachbarn aufgrund der Peptidbindung nicht

in Kontakt stehen können, da diese einen maximalen Abstand vorgibt, um den sich die Nachbarn im Vergleich zu dem Paar näher kommen können, für das der Abstand bestimmt wurde. Die Berechnungszeit für die Distanzkontaktrelationen eines Proteins ist neben der Anzahl der Aminosäurereste auch von der Anzahl der betrachteten Atome und vom maximalen Abstand abhängig, bis zu dem ein Kontakt gezählt wird. Dieser Maximalabstand und die verwendeten Atome (zum Beispiel C_α oder C_β) kann von Potential zu Potential unterschiedlich sein. Daher werden die Distanzkontaktrelationen der Proteine sowohl für die Ableitung der Potentiale als auch in Faltungserkennungsexperimenten zur Laufzeit neu berechnet.
In dem Distanzkontaktrelationsmodell kann nicht garantiert werden, daß die in Relation stehenden Aminosäurepositionen in der Struktur auch direkt benachbart sind. Im Gegenteil ist davon auszugehen, daß zum Beispiel bei einem Abstand von 20Å zwischen zwei Aminosäuren Teile anderer Aminosäuren liegen. Ziel der an der GMD entwickelten *Voronoi*–Potentiale [382] ist die Bewertung von Relationen in Proteinen, die aus echten geometrischen Nachbarschaften bestehen. Die *Voronoizerlegung* [355] einer Punktmenge im Raum liefert die dafür geeignete Datenstruktur für derartige echte Nachbarschaften. Dazu werden die Atome eines Proteins als Punkte im $\mathbb{R}^3$ betrachtet.

Definition 5.4 (Voronoizerlegung)
Seien $\vec{p_i}$ und $\vec{p_j}$ zwei Punkte im $\mathbb{R}^3$. Dann zerlegt die Ebene

$$E(\vec{p_i}, \vec{p_j}) = \{\vec{x} \in \mathbb{R}^3 : \|\vec{x}, \vec{p_i}\| = \|\vec{x}, \vec{p_j}\|\}$$

den $\mathbb{R}^3$ in zwei Halbräume $H_{E(\vec{p_i},\vec{p_j})}(\vec{p_i})$ und $H_{E(\vec{p_i},\vec{p_j})}(\vec{p_j})$ mit $\vec{p_i} \in H_{E(\vec{p_i},\vec{p_j})}(\vec{p_i})$ und $\vec{p_j} \in H_{E(\vec{p_i},\vec{p_j})}(\vec{p_j})$.
Seien $PB \subset \mathbb{R}^3$ die Menge aller Atomkoordinaten eines Proteins. Dann ist die Voronoizelle *$V(\vec{p_i})$ eines Punktes $\vec{p_i} \in PB$ definiert als der Schnitt der Halbräume mit den anderen Punkten aus PB:*

$$V(\vec{p_i}) = \bigcap_{\vec{p_j} \in B, \vec{p_j} \neq \vec{p_i}} H_{E(\vec{p_i},\vec{p_j})}(\vec{p_i})$$

Die Voronoizelle *$V(b)$ einer Aminosäure $b \in B$ ist die Vereinigung der Voronoizellen der Atome von b:*

$$V(b) = \bigcup_{\vec{p_i} \in b} V(\vec{p_i})$$

Die Voronoikontaktfläche *$VKF(a, b)$ zweier Aminosäuren $a, b \in B$ ist definiert als die Summe der Größen der gemeinsamen Flächen der zugehörigen Voronoizellen $V(a)$ und $V(b)$.*
Die Voronoizerlegung *eines Proteins ist die Vereinigung der Voronoizellen seiner Aminosäuren.*

Die Voronoizelle der Atomkoordinaten zerlegen den Raum in konvexe Polytope. Die Zellen der Atompunkte, die auf der konvexen Hülle des Proteins liegen, sind zunächst einmal unbeschränkt. Damit sind auch die Voronoikontaktfläche zweier auf der konvexen Hülle benachbarter Aminosäuren unbegrenzt.
Um auch die Voronoizellen an der Proteinoberfläche eine der Größe der Aminosäure proportionale Ausdehnung zu geben, werden zusätzlich zu den Atompunkten sogenannte *Gitterpunkte* eingefügt, die das Protein gewissermaßen mit einer Lösungsmittelschicht umgeben.

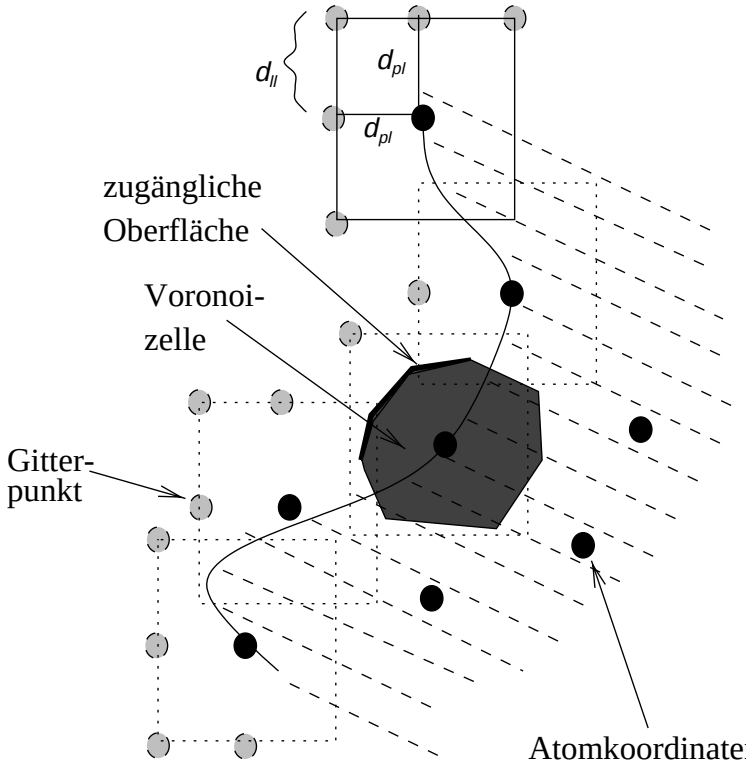


Abbildung 5.1: Voronoizerlegung eines Proteins mit Atomkoordinaten (schwarze Kugeln) und Gitterpunkten (graue Kugeln).

Abbildung 5.1 zeigt einen Ausschnitt an der Oberfläche eines Proteins. Die Gitterpunkte werden auf einem Würfel mit Kantenlänge $2*d_{pl}$ um jede Atomkoordinate, die den Mittelpunkt des Würfels bildet, nach folgenden Vorschriften generiert:

- Ein Gitterpunkt ist von jedem Atompunkt mindestens d_{pl}Å entfernt.
- Ein Gitterpunkt ist von jedem anderen Gitterpunkt mindestens d_{ll}Å entfernt.

Durch die Einführung der Gitterpunkte wird das Volumen der Aminosäuren begrenzt, die die Oberfläche der konvexen Hülle des Proteins berühren. Damit werden können die Voronoikontaktflächen von Aminosäuren an der Proteinoberfläche

einheitlich zu denen von vergrabenen Aminosäureresten berechnet werden. Außerdem kann die Summe $SVF(b)$ der Flächengrößen, die die Voronoizelle $V(b)$ einer Aminosäure b mit Gitterpunkten gemeinsam hat, als Maß für die Lösungsmittelzugänglichkeit einer Aminosäure b verwendet werden.
Die Kalibrierung der Gitterparameter d_{pl} und d_{ll} erfolgt anhand bereits aus der Literatur bekannter Größen für die zugängliche Oberfläche und das Volumen einer Aminosäure. Als Referenzgröße für die zugängliche Oberfläche $SVF(b)$ einer Aminosäure b dient die *Connolly*–Oberfläche [66] aus `DSSP` [173]. Als Vergleich für das typische Voronoivolumen einer Aminosäure werden die mittleren Volumen von Aminosäuren eines Typs aus der Literatur [133, 277] verwendet, die in beiden Referenzen auf der Basis von Voronoizerlegungen der vollständig im Proteininneren vergrabenen Aminosäuren berechnet wurden.
Eine ausführlichere Beschreibung der Bestimmung der optimalen Gitterparameter findet sich in [366, 367, 382]. Dabei wurde bei der Kalibrierung berücksichtigt, daß nach Gerstein et al. [114] das Volumen, das von Atomen an der Proteinoberfläche eingenommen wird, etwa 6% größer als das von vergrabenen Atomen ist. Die gegenwärtige Einstellung ist 3Å sowohl für d_{pl} als auch d_{ll}. Dies entspricht intuitiv in etwa dem Durchmesser eines Wassermoleküls.
Für die Voronoikontaktdefinition (siehe 5.5) stellen unbegrenzte Flächen an der Problemoberfläche nur dann ein Problem dar, wenn einem Voronoikontakt eine sinnvolle Kontaktflächengröße zugewiesen werden soll. Für die reine Nachbarschaftsrelation ist nur die Existenz einer gemeinsamen Fläche von Bedeutung.

Definition 5.5 (Voronoikontaktrelation)
Aminosäurereste $[b] \in B$ *stehen in einen* Voronoikontaktrelation $\sim_v$ *genau dann, wenn gilt:*

$$\forall (b_i, b_j) \in [b] \ : \ VKF(b_i, b_j) > 0$$

Der Abstand $max_{(b_i,b_j)\in[b]} \|b_i, b_j\|$ *ist ein Attribut des Kontaktes* $\sim_v [b]$.
Ein weiteres Attribut eines Voronoikontaktes ist seine Gesamtkontaktfläche

$$VKF[b] = \sum_{(b_i,b_j)\in[b]} VKF(b_i, b_j).$$

Die Berechnung der Voronoikontaktrelationen eines Proteins erfolgt über die Berechnung der konvexen Hülle von Punkten im $\mathbb{R}^4$ mit Hilfe des `Quickhull`–Algorithmus [22]. Die Laufzeit zur Berechnung der konvexen Hülle im $\mathbb{R}^4$ und damit der Voronoizerlegung im $\mathbb{R}^3$ erfolgt in asymptotischer Zeit $O(n^2)$ [87, 370], wobei n die Anzahl der Eingabepunkte also Atompunkte plus Gitterpunkte ist. Zur Transformation des Problems der Berechnung der Voronoizerlegung von Punkten im $\mathbb{R}^d$ auf die Berechnung einer konvexen Hülle im $\mathbb{R}^{d+1}$ wird die Methode von Edelsbrunner und Seidel [88] verwendet (siehe auch [265, 366, 370]).
Wie die Voronoi–Datenstruktur effizient zur Berechnung der Voronoikontaktrelation und der zugehörigen Kontaktattribute verwendet werden kann, ist in der

Diplomarbeit von Marko Woehler [366] im Detail beschrieben. Die Berechnung der Voronoikontaktrelationen eines Proteins aus den in der PDB–Datenbank abgelegten Proteinkoordinaten ist mit den Zwischenschritten insgesamt etwas zeitaufwendiger als die Berechnung der Distanzkontaktrelation.
Da die Voronoikontaktrelationen außerdem den Vorteil haben, unabhängig von dem in einem empirischen Potential maximal bewerteten Abstand zweier Aminosäurereste zu sein, wurden die Voronoikontaktrelationen vorberechnet und in einer Datenbank abgelegt.
Damit ergibt sich folgende formale Definition des Relationsgraphen G_B einer Struktur B:

Definition 5.6 (Relationsgraph)
Sei $\mathcal{I}$ die Menge der in den Definitionen 5.2 bis 5.5 definierten (Kontakt–)Relationen. Der Relationsgraph G_B^I für eine Proteinstruktur B bezüglich einer Menge $I \subset \mathcal{I}$ ist dann definiert als ein Quadrupel (V_B, E_B, λ, μ) der Knotenmenge V_B, der Kantenmenge V_B, den Knotenattributen λ und den Kantenattributen μ mit:

- *V_B Menge der Aminosäurereste der Struktur,*
- *$E_B = \{[b] \in V_B \ : \sim_i ([b])$, wobei $\sim_i \in I\}$,*
- *Knotenattributen $\lambda \ : \ V_p \longrightarrow \Sigma$,*
- *Kantenattributen $\mu \ : \ E_B \longrightarrow \mathcal{I} \times \mathbb{Z}^k \times \mathbb{R}^l$.*

Attribute, die Eigenschaften einzelner Positionen in der Struktur beschreiben, werden dabei als einstellige Relationen kodiert. Ein Beispiel für eine einstellige Relation ist die lösungsmittelzugängliche Oberfläche eines Aminosäurerestes. Relationen an denen mehr als zwei Positionen beteiligt werden als sogenannte *Hyperkanten* kodiert. Bis auf wenige Ausnahmen [119] werden in der Faltungsvorhersage nur Relationen mit Grad kleiner als drei verwendet.
Die Knotenattribute λ ordnen den Knoten unter anderem den Aminosäuretyp aus der nativen Struktur zu. Die Kantenattribute μ dienen der Kodierung der zusätzlichen Attribute der der jeweiligen Kante entsprechenden Relation. In Abhängigkeit von Typ des Attributs und der Relation werden sie als ganze oder reelle Zahlen an den Kanten des Relationsgraphen gespeichert. Typische Kantenattribute sind der euklidischen Abstand (ED) der durch die Kante verbundenen Knoten, der Abstand (SD) der Knoten in der zugrundeliegenden Proteinsequenz oder auch die Größe der Kontaktfläche (KF) einer Kontaktrelation.
Ein Ereignis im Sinne der für die Potentialableitung benutzten Terminologie ist also ein Element der Kantenmenge des Relationsgraphen mit den zugehörigen Kantenattributen. Um auch bei Verfeinerungen der Relationen zum Beispiel durch ED–Attribute mit diskreten Ereignissen zu rechnen, werden Abstandsattribute

und andere Attribute, die von ihrer Art her kontinuierlich sind, durch Unterteilung in Intervalle diskretisiert. Somit können sowohl bei der statistischen Ableitung von Potentialen diskrete Ereignisse gezählt als auch bei ihrer Anwendung diskrete Ereignisse bewertet werden.
Der Relationsgraph G_B stellt somit eine formale Beschreibung einer Proteinstruktur B dar, anhand der im folgenden die verschiedene Ein– und Mehrkörperpotentiale beschrieben werden. Dabei werden insbesondere Unterschiede der aus der Literatur bekannten Potentiale zu den an der GMD entwickelten Potentialen diskutiert, die auf der Voronoizerlegung von Proteinen basieren.

5.2.3 Einkörperpotentiale

Unter dem Begriff Einkörperpotentiale werden hier die Potentiale gefaßt, die das Auftreten einer Aminosäure an einer bestimmten Strukturposition nur in Abhängigkeit von dem Typ der Aminosäure und einer Beschreibung der Strukturposition abhängig machen. Während die einfachste Variante von Einkörperpotentialen – die sogenannten Hydrophobizitätspotentiale – für die Präferenz einer Aminosäure kodieren, sich an einer dem Lösungsmittel zugänglichen Position oder an einer im Proteinkern vergrabenen Position der Struktur zu befinden, werden in den sogenannten Profilpotentialen Präferenzen kodiert, die auf der einer komplexeren Beschreibung der Strukturposition kodieren.
Wie bereits in Abschnitt 4.2.2 haben alle Einkörperpotentiale unabhängig von der jeweiligen Beschreibung einer Strukturposition den großen Vorteil, daß das bezüglich der zugehörigen Kostenfunktion optimale Sequenzstrukturalignment mit den Methoden der dynamischen Programmierung berechnet werden kann.

5.2.3.1 Hydrophobizitätspotentiale

Hydrophobizität ist bekanntlich eine der treibenden Kräfte der Proteinfaltung. Daher wird dieser Term auch in empirischen Potentialen zur Faltungserkennung benutzt. Im folgenden wird ein Auswahl verschiedener empirischer Hydrophobizitätspotentiale vorgestellt:

Vergraben oder zugänglich: Godzik und Skolnik [119] leiten ein aminosäuretypabhängiges Hydrophobizitätspotential nach dem in Abschnitt 5.2.1 beschriebenen Verfahren ab. Als Ereignis definieren sie das Auftreten eines Aminosäuretyps an der Proteinoberfläche.

Paarkontaktabhängig: Casari und Sippl [52] leiten Hydrophobizitätspotential über Hauptkomponentenanalyse aus dem Paarinteraktionspotential von Sippl (siehe Abschnitt 5.2.4) her.

Bryant und Lawrence [48] berechnen den Hydrophobizitätsanteil des Pseudoenergiepotentials ebenfalls durch Komponentenzerlegung des Paarinteraktionspotentials in paarabhängige und paarunabhängige Komponenten.

Hydrophobe Kontakte: Hang *et al.* [154, 155] schlagen ein sehr einfaches Hydrophobizitätsmaß vor, das aus einem Hydrophobizitätsterm und einem Term für die Eigenschaft einer Aminosäure besteht, im Proteininneren vergraben zu sein. Der Hydrophobizitätsterm für eine hydrophobe Aminosäure an Position i errechnet sich aus der Anzahl der nichtpolaren Aminosäuren, die einen Kontakt (Abstand der virtuellen C_β–Atome ≤ 7.3Å) zu Position i haben. Der zweite Term errechnet sich einfach aus der Anzahl Reste in einem 10Å Radius um i. Die Erkennungsrate im Selbsterkennungstest geben die Autoren mit 85% an.

Abstand zum Zentroid: DeBoolt und Skolnik [75] definieren ein auf Atomebene berechnetes Einkörperpotential über den Abstand eines Atoms vom Zentroid des Proteins B im Verhältnis zum Gyrationsradius (abgeschätzt durch $2.2 * |B|^{0.38}$). Ein gezähltes Ereignis besteht darin das ein Atom in einer bestimmten Umgebungshülle um den Proteinzentroid vorkommt. Das Potential bewertet somit indirekt die Präferenz eines Atoms vergraben zu sein.

Oberflächenanteil : In [366] wird die mit Gitterpunkten der Voronoizerlegung gemeinsame Fläche eines Aminosäurerestes in Verhältnis zur Gesamtoberfläche der Voronoizelle des Restes gesetzt. Durch Diskretisierung dieser Anteile in 10 Intervalle zwischen 0 und 1.0 wird die in Abschnitt 5.2.1 beschriebene Potentialableitungmethodik anwendbar. Jones *et al.* [168] leiten ein Hydrophobizitätspotential auf Basis der Connolly–Oberfläche [66] ab, indem sie für eine Aminosäure vom Typ `X` die in der Trainingsmenge gefundenen Connolly–Oberflächen in Bezug zur Oberfläche dieser Aminosäure im Pentapeptid `GGXGG` setzen.

Während Hydrophobizitätspotentiale im wesentlichen die Präferenz von Aminosäureresten kodieren, im Proteininneren vergraben zu sein, werden in dem im folgenden Abschnitt beschriebenen Profilpotentialen Präferenzen von Aminosäureresten für weitere Eigenschaften einer strukturellen Umgebung kodiert.

5.2.3.2 Profilpotentiale

Profilepotentiale sind Bewertungssysteme, die eine Präferenz eines Aminosäurerestes für eine bestimmte Strukturumgebung kodieren. Einer der ersten Ansätze zur Faltungserkennung durch Profilpotentiale wurde 1991 von Bowie, Lüthy und Eisenberg [39] vorgeschlagen. Eine Strukturumgebung wird in diesem Ansatz durch

- die Oberfläche eines Restes, die im Protein vergraben ist,
- den Oberflächenanteil eines Restes, der in Kontakt zu polaren Atomen oder Wasser steht, und
- die lokale Sekundärstruktur beschrieben.

Der Vorteil einer derartigen Vorgehensweise ist, daß eine Struktur einfach als ein Profil, dessen Länge durch die Länge der Struktur und dessen Tiefe durch die Anzahl der für die Beschreibung struktureller Umgebungen verwendeten Attribute beschrieben werden kann. Der wichtigste Vorteil ist aber, daß ein Alignment einer Sequenz gegen diese Repräsentation einer Struktur optimal mit den aus dem Sequenzalignment bekannten Methoden berechnet werden kann. In [93, 209, 210, 221, 267, 356, 357, 374] werden andere Beschreibungen von Strukturumgebungen vorgestellt.
Ouzonis *et al.* [266] verwenden eine Strukturbeschreibung, die sogenannten *contact interface vectors*, in der für jede Strukturposition zum Beispiel die eigene Sekundärstruktur, die Anzahl und Stärke von Kontakten zu Resten in der gleichen und anderen Sekundärstrukturen und zum Lösungsmittel kodiert werden. Im nächsten Schritt werden dann, ähnlich wie in Abschnitte 5.2.1 beschrieben, für jeden Aminosäuretyp Präferenzen abgeleitet, sich in einer so beschriebenen Strukturumgebung zu befinden, beziehungsweise mit dem zugehörigen *contact interface vector* aliniert zu werden.
Das an der GMD entwickelte Kontaktkapazitätspotential (CCP) [5], welches als Kostenfunktionsbestandteil in dem schnellen Faltungserkennungsverfahren `123D` [5] verwendet wird, weist in eine ähnliche Richtung wie der zuvor beschriebene Ansatz und soll etwas ausführlicher beschrieben werden, da das Kontaktkapazitätspotential auch in der RDP–Methode verwendet wird. Dieses Kontaktkapazitätspotential charakterisiert einen Aminosäurerest anhand seiner Möglichkeit eine bestimmte Anzahl von Kontakten zu anderen Resten in einer Proteinstruktur auszubilden. Das C_β–Atom in Abbildung 5.2 zum Beispiel befindet sich in einer Schleife und hat 2 Kontakte zu nah in der Sequenz benachbarten Positionen und zu weiteren 5 in der Sequenzreihenfolge weiter entfernten Positionen. Entsprechend kodiert die einfachste Variante CCP eines Kontaktkapazitätspotentials, die Präferenz einer Aminosäure, Bestandteil einer bestimmten Sekundärstruktur zu sein und dabei k lokale und l globale Kontakte zu haben. Lokale Kontakte sind Kontakte zu maximal 6 Peptidbindungen entfernten Positionen, globale Kontakte alle anderen. Der Aminosäuretyp des Restes zu dem der Kontakt ausgebildet wird, wird nicht berücksichtigt. Dies ist der Grund, warum CCP–Potentiale in Profilen kodiert werden können und das zugehörige Alignmentproblem effizient gelöst werden kann.
Die Zählung der Vorkommen eines Aminosäuretyps in einer bestimmten Kontaktumgebung werden, wie beschrieben, über das inverse Boltzmanngesetz in Pseudoenergie beziehungsweise Präferenzen umgewandelt. In [5] werden weitere Verfeinerungen der Kontaktkapazitätspotentiale vorgeschlagen:

Bedingtes *CCP* (*CCCP*) : Die Präferenz für eine bestimmte Anzahl globaler Kontakte eines Restes wird unter der Bedingung bestimmt, daß eine bestimmte Anzahl lokaler Kontakte vorliegt.

Abstandsabhängiges *CCP* (*DCCP*) : Die Kontakte werden nach dem von

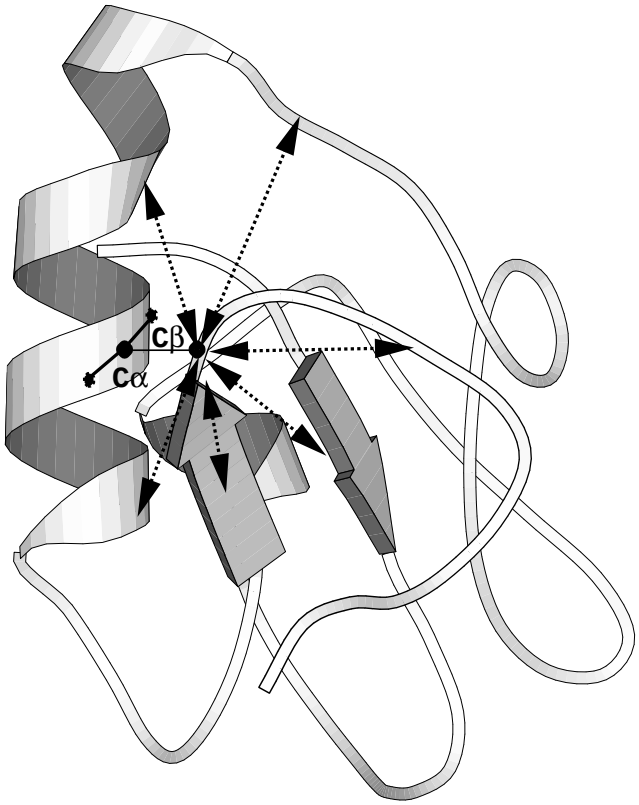


Abbildung 5.2: Die Kontaktkapazitätpotentiale [5] bewerten die Präferenz einer Aminosäure in einer bestimmten Kontaktumgebung aufzutreten. Die Kontaktumgebung einer Strukturposition C_β wird zum Beispiel durch die Anzahl der Kontakte zu anderen Strukturpositionen und die Sekundärstruktur beschrieben.

Bryant und Lawrence [48] vorgeschlagenen Schema zusätzlich nach dem euklidischen C_β–Abständen und der Trennung der Reste in der Sequenz klassifiziert.

Winkelabhängiges *CCP* (*ACCP*) : Die den betrachteten Rest umgebende Raumkugel wird in 6 Winkelsegmente eingeteilt. Die Kontakte werden entsprechend ihrer Zuordnung zu einem dieser Segmente klassifiziert.

In der RDP–Methode wird bisher nur die einfache Version *CCP* verwendet. Eine Erweiterung auf die anderen Potentialtypen ist jedoch ohne Änderung der RDP–Methode möglich. Da jedoch das einfachste Potential auch in der `123D`–Methode bislang die besten Ergebnisse liefert, wurde bisher darauf verzichtet. Auf den Einsatz der *CCP*–Potentiale in der RDP–Methode wird in Kapitel 7 eingegangen.

5.2.4 Zwei– und Mehrkörperpotentiale

Zwei– und Mehrkörperpotentiale bewerten wie der Name bereits sagt Ereignisse e, an denen zwei oder mehr Aminosäuren eines Proteins beteiligt sind.
Die ersten empirischen Zweikörperpotentiale wurden 1985 von Miyazawa und Jernigan [232] vorgestellt. Sippl erweiterte 1990 in seiner Beschreibung der *Potentials of Mean Force* [319] diese empirischen Kontaktpotentiale zu distanzabhängigen Paarkontaktpotentialen. In der Literatur wurden seitdem von verschiedenen Arbeitsgruppen zahlreiche, leicht von der Ursprungsversion abweichende, empirische Potentiale vorgestellt. Der folgenden Abschnitt gibt einen Überblick über die Vielfalt der in der Literatur vorgeschlagenen Variationen.

- Miyazawa und Jernigan [232] transformieren die Seitenkettenzentren auf ein Gitter und belegen nicht besetzte Gitterpunkte mit Lösungsmittel. Gitterpunkte in 6.5Å definieren sie als Kontakte, die nach Lösungsmittel und Aminosäuretyp getrennt gezählt werden. Die Zählungen werden über die quasi–chemische Approximation in Energiewerte übersetzt. Sie verwenden die so abgeleiteten Potentiale für Simulationen und Faltungserkennung [233]. Die wesentlichen Unterschiede zu anderen Potentialen bestehen in der expliziten Berücksichtigung des Lösungsmittels und darin, daß sie nicht distanzabhängig sind.
- In seiner ersten Arbeit zu empirischen Potentialen [319] beschreibt Sippl ein aminosäuretypabhängiges, empirisches Paarinteraktionspotential, dessen Ableitung auf dem in Abschnitt 5.2.1 beschriebenen Grundprinzip beruht. Als erstes Attribut eines Kontaktes verwendet er den euklidischen Abstand (ED) der C_α–Atome zweier Aminosäuren im Bereich zwischen etwa 2.5 und 22Å), wobei er diese in etwa zwanzig Intervalle unterteilt. Als zweites Attribut schlägt er den Abstand der Reste in der Aminosäuresequenz (SD) vor, wobei zunächst nur Aminosäurereste im Sequenzabstand zwischen 2 und 6 in die Betrachtung einbezogen werden (sogenannte *kurzreichweitige* Kontakte). Da in diesem Modell die Anzahl der Vorkommnisse für einige Ereignisse zu klein ist, verwendet Sippl einen Korrekturterm.
- Für Faltungserkennungsexperimente [139, 325] werden die Abstände der C_β–Atome und außerdem auch Kontakte zwischen weiter in der Sequenz voneinander entfernten Resten verwendet. Wie für die ED–Attribute werden dazu auch die SD–Attribute für die mittel– und langreichweitigen Kontakte in Intervalle aufgeteilt.
- Jones *et al.* nutzen ein Potential, wie es von Sippl eingeführt wurde, in ihrem Programm `Threader` für die Berechnung von Sequenzstrukturalignments und Faltungserkennungsexperimente [168].
- Bryant und Lawrence [48] verwenden bis auf die Wahl des Referenzzustandes (siehe dazu Abschnitt 5.2.1) eine ähnliche Vorgehensweise wie Sippl. Sie

definieren Kontakte zwischen virtuellen C_β–Atomen (siehe oben) im euklidischen Abstand zwischen 5 und 10Å in Intervallschritten von 1Å. Außerdem beschränken sie das Potential auf Kontakte mit Sequenzabstand größer als 5. Daher unterscheiden sie Kontakte auch nicht nach dem Sequenzabstand der Kontaktpartner, sondern anstelle dessen danach, ob die Kontakte vom Rückgrad der Peptidkette oder von den Seitenketten ausgebildet werden.

- Bauer und Beyer [24] machen in ihrer Potentialableitung den Korrekturterm aus dem Sippl–Ansatz Ansatz überflüssig, indem sie für ein Aminosäurepaar (a, b) beobachtete Ereignisse gewichtet über eine Aminosäureaustauschmatrix auch als Ereignisse für alle anderen Paaren (x, b) registrieren.
- Levitt *et al.* verwenden Variationen der bereits vorgestellten Potentiale [268, 269].
- Kocher *al.* [184] erweitern die typischen Sippl–Potentiale um einen Rückgratdihedralwinkelterm und einen Lösungsmittelterm.
- Reva *et al.* [280] verwenden ebenfalls eine Variation des Sippl–Potentials.
- Maiorov und Crippen [212] repräsentieren die Seitenkette ebenfalls durch das C_β–Atom (Kontaktabstand maximal 9Å), verwenden aber zusätzlich auch Kontakte zu Rückgratatomen (Kontaktabstand maximal 5Å). Außerdem unterscheiden sie Kontakte zwischen Rückgratatomen, zwischen Seitenketten– und Rückgratatomen und zwischen Seitenkettenatomen. Die Bestimmung der Werte des empirischen Potentials erfolgt nicht durch Verwendung des inversen Boltzmanngesetzes sondern durch Lösung eines großen Ungleichungssystems. Die Ungleichungen verlangen, daß jede native Struktur eine geringere Energie erhält als jede Alternative in der Trainingsmenge. Die Lösung des Ungleichungssystems erfolgt durch schrittweise Vergrößerung der Trainingsmenge und damit des Ungleichungssystems. Bei der Anwendung der Potentiale heben sie die Diskretisierung der Kontakte auf und bewerten Kontakte kontinuierlich in Abhängigkeit von der Distanz.
- Godzik und Skolnik [119] leiten ein aminosäuretypabhängiges Mehrkörperpotential nach der in Abschnitt 5.2.1 beschriebenen Vorgehensweise ab. Als aminosäuretypabhängiges Ereignis definieren sie eine Distanzkontaktrelation von zwei beziehungsweise drei Aminosäureresten. Zwei beziehungsweise drei Aminosäurereste stehen in Kontakt, wenn für jedes Paar zwei schwere Seitenkettenatome existieren, deren Abstand kleiner als 5Å [118]. Dieses Modell unterscheidet sich in zwei Punkten von dem von Sippl und Bryant verwendeten Modell:
  - Es sind auch dreistellige Relationen möglich.
  - Es wird wie auch in [232] nur die Existenz eines Kontaktes und keine weitere Eigenschaften (wie zum Beispiel die Distanz) berücksichtigt.

- In [323, 324] leiten Sippl *et al.* ein empirisches Potential für die Bewertung von Wasserstoffbrücken aus den beobachteten Distanzabständen zwischen den Carboxylsauerstoffatomen und den Amidgruppen nicht in der Sequenz benachbarter Peptideinheiten ab.
- DeBoolt und Skolnik [75] berechnen ein distanzbasiertes empirisches Potential auf Atomebene und verwenden es zur Diskriminierung zwischen nativen Faltungen und Mißfaltungen. In [225] werden sogar noch die einzelnen Atome gleichen Typs nach weiteren Kriterien unterschieden.

In der RDP–Methode ist es möglich, verschiedene Arten von Einkörper– und Mehrkörperpotentialen zur Berechnung und Bewertung von Sequenzstrukturalignments zu verwenden. Zur Zeit werden Ein– und Zweikörperpotentiale eingesetzt.
Dabei werden im wesentlichen zwei unterschiedliche Typen von Potentialen verwendet. Potentiale des ersten Typs sind Variationen der von Sippl beziehungsweise Bryant und Lawrence vorgeschlagenen Potentiale, also Potentiale, die auf den Distanzkontaktrelationen (siehe Definition 5.3, C_β–Distanzen) beruhen.
In diesen Potentialen werden Kontakte neben den Aminosäuretypen der Kontaktpartner durch ihren Abstand im Raum (ED) und in der Sequenz (SD) unterschieden. Diese Potentiale wurden in [366] hinsichtlich ihrer Erkennungsleistung im Selbsterkennungstest optimiert. Die Erkennungsleistung in diesem Test steigt mit dem Detailgrad des Potentials, der über die Anzahl der verwendeten Intervalle für den Sequenz– und räumlichen Abstand der Kontaktpartner beeinflußt wird. Die Anzahl der verwendeten Intervalle ist im folgenden im Namen des Potentials kodiert. So bedeutet `ED3SD6` zum Beispiel, daß die Kontakte nach ihrem Abstand im Raum zwischen 3 und 15Å in sechs 2Å–Intervalle und in sechs Sequenzabstände ($2, 3, 4, 5, 6$ und ≤ 7 bis ∞) eingeteilt sind.
In einer homologen Struktur ist zwar das grobe Faltungsgerüst erhalten, aber einzelne Distanzen können sich durch mehrere Angstrœm unterscheiden. In [131] werden diese Unterschiede diskutiert. Bei der Optimierung der Erkennungsleistung wurde daher versucht, den Detailgrad der Potentiale, also die Anzahl der Intervalle, bei optimaler Erkennung der nativen Strukturen möglichst gering zu halten. Für weitere Details, wie die Wahl der Trainingsmenge, des Referenzzustandes und die Korrektur für kleine Datenmengen, sei hier auf die Diplomarbeit von Marko Woehler [366] verwiesen.
Potentiale des zweiten Typs basieren auf dem Voronoikontaktrelationen (siehe Definition 5.5, C_β–Distanzen). Potentiale dieses Typs werden im folgenden durch das Präfix `VCM` gekennzeichnet. Dieses Modell hat neben der besseren Erkennungsleistung (siehe Abschnitt 5.2.5) den Vorteil, daß auch die Kontaktfläche KF eines Kontaktes in seine Bewertung einbezogen werden kann. Dazu bietet die RDP–Methode zum Sequenzstrukturalignment zwei Möglichkeiten:

- Gewichtung eines Kontaktes anhand der Größe der zugehörigen Kontaktfläche, eventuell mit Gewicht 0 für sehr kleine Kontaktflächen.

- Verwendung der Kontaktfläche KF anstelle des Sequenzabstandes SD oder des euklidischen Abstandes ED als Kriterium für die Attributierung eines Kontaktes.

Die Erkennungsleistung von den Kontaktflächenpotentiale im Selbsterkennungstest zeigt Tabelle 5.1 [367, 382].
Wie bereits erwähnt, bestehen Voronoikontaktrelationen nur zwischen im Raum direkt benachbarten Aminosäuren. In wieweit dadurch langreichweitige physikalische Wechselwirkungen, wie zum Beispiel elektrostatische Wechselwirkungen, nicht hinreichend durch ein `VCM`–Potential erfaßt werden, bleibt noch offen.
Bei der Berechnung von Sequenzstrukturalignments mit der RDP–Methode stellen die Bewertung von (Teil–)Alignments und die Kodierung von Paarinteraktionen zentrale Berechnungsschritte dar. Der für diese Schritte benötigte Zeitaufwand steigt mit der Anzahl der auszuwertenden Kontakte. Daher wirkt sich die Beschränkung auf wirkliche Nachbarschaften bei Voronoikontaktrelationen als Nebeneffekt positiv auf die Laufzeit von Sequenzstrukturalignmentberechnungen mit der RDP–Methode aus.

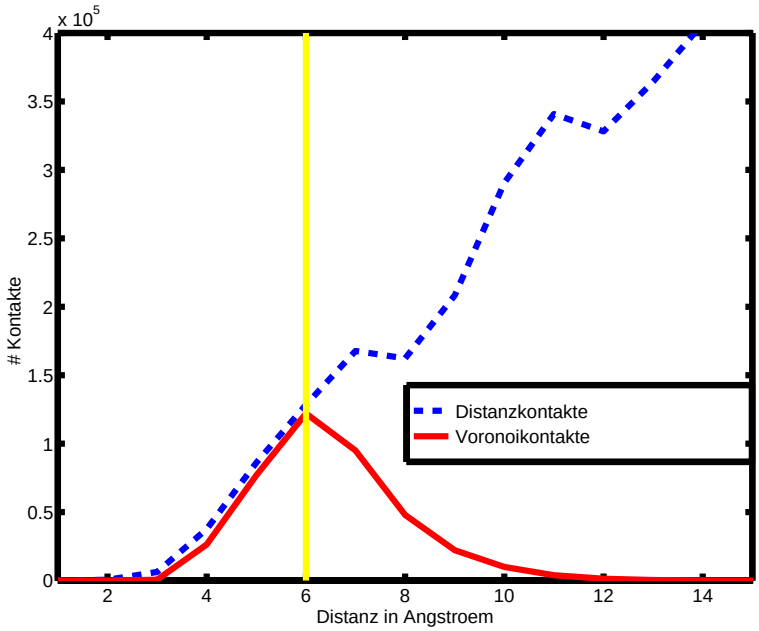


Abbildung 5.3: Anzahl Voronoikontakte (in rot) versus Distanzkontakte (in blau) für eine repräsentative Proteinmenge (hobohm_96_25, 488 Proteine) [366].

Abbildung 5.3 zeigt für eine repräsentative Proteinmenge (`hobohm_96_25`, 488 Proteine) die Anzahl der Kontakte in Abhängigkeit von der Distanz für die Distanzkontaktrelationen und die Voronoikontaktrelationen. Bis 6Å verhalten sich die Relationen nahezu identisch. Doch für größere Distanzen nimmt die Anzahl

der Kontakte bei Distanzkontaktrelationen weiter zu, während sie für Voronoikontaktrelationen recht schnell abnimmt.
Der folgende Abschnitt gibt eine Einordnung des Nutzens empirischer Potentiale und setzt die Erkennungsleistung der neuartigen auf den Voronoikontaktrelationen basierten Potentiale in Bezug zu aus der Literatur für andere Potentiale entnommenen Erkennungsraten.

5.2.5 Bewertung empirischer Potentiale

Empirische Potentiale können immer nur eine Approximation der freie Energie ΔG sein. Dies bedeutet aber auch, daß die native Struktur, die allgemeinhin als die Struktur minimaler freier Energie angesehen wird, nicht notwendigerweise die minimale Energie gemäß eines empirischen Potentials haben wird, sondern daß Strukturen mit niedrigerer Energie möglich sind. Es gibt jedoch Hinweise dafür, daß die Anzahl der möglichen Konformationen einer Peptidkette exponentiell mit der Energie steigt [98]. Dies erschwert die Verwendung empirischer Potentiale für Moleküldynamiksimulationen und Energieminimierungen, da am Ende nicht eine Struktur sondern ein Ensemble von Strukturen steht, aus dem die zuverlässigen Teile extrahiert werden müssen [316]. Dennoch werden spezielle Anpassungen von empirischer Paarinteraktionspotentialen auch für Simulationen der Proteinfaltung in Gittern verwendet [283]. In einer Übersicht vergleicht Moult [238] die klassischen Kraftfelder der molekularen Mechanik mit den empirischen, aus Datenbanken abgeleiteten Potentialen und kommt zu dem Schluß, daß beide Ansätze noch viele offene Fragen und Probleme beinhalten.
Während die empirischen Potentiale für *Ab initio*–Simulationen noch zu ungenau zu sein scheinen, sind sie für die Faltungserkennung sehr hilfreich [74]. In [320, 322] gibt Sippl eine Übersicht über den Nutzen empirischer Potentiale in der Proteinstrukturvorhersage.
Bei der Faltungserkennung ist jedoch das Ziel, zwischen Faltungsklassen oder zumindest deutlich unterschiedlichen Faltungsmodellen für eine gegebene Sequenz zu unterscheiden. Für die Faltungserkennung kommt es daher darauf an, eine Potentialfunktion zu finden, die eine Balance zwischen Detailgrad und Abstraktion hat. Das Abstraktionslevel muß so hoch sein, daß die Erkennung einer homologen Struktur nicht an der Ersetzung von etwa 80% der Aminosäuren scheitert. Auf der anderen Seite muß der Detailgrad aber auch so groß sein, daß eine Unterscheidung zwischen homologen und nicht homologen Strukturen möglich ist. Dazu ist jedoch die Berechnung einer korrekten Abbildung der Sequenz in die homologe Struktur notwendig. Das wiederum bedeutet, daß es mit Hilfe der Potentialfunktion möglich sein muß, zwischen guten und schlechten Sequenzstrukturalignments zu unterscheiden.
Der einfachste Test, dem Potentiale hinsichtlich ihrer Erkennungsleistung unterworfen werden können, ist der sogennannte Selbsterkennungstest [139]. Dieser Test wird in der Literatur allgemein verwendet, um die Leistungsfähigkeit ei-

nes vorgeschlagenen empirischen Potentials zu testen. In diesem Test wird eine Sequenz ohne Insertionen und Deletionen durch eine repräsentative Menge von Strukturen geschoben, in der sich auch die zu der Sequenz gehörige native Struktur befindet. Jede der so erhaltenen Einbettungen wird anhand des Potentials bewertet. Ein Sequenzstrukturpaar gilt als erkannt, wenn die Identitätsabbildung der Sequenz B in die native Struktur B die niedrigste Pseudoenergie erhält, also an Position 1 in der Rangliste R_B steht.

	Hendlich *et al.* [139]	Bryant & Lawrence [48] [1]	DeWitte & Shakhnovich [77]	Kocher *et al.* [184]	Bauer & Beyer [24]	Huang *et al.* [154]	Reva *et al.* [280]
#	101	161	86	74	167	195	102
$\mathcal{R}[\infty]$ [%]							
Original	66	94	-	73	94	85	98
ED3SD6	93	88	100	100	96	98	100
ED6KF8	95	91	100	100	98	98	100
KF2SD6	92	90	97	98	97	98	100

Tabelle 5.1: Selbsterkennungstest: Vergleich verschiedener Voronoipotentiale gegen Potentiale aus der Literatur. ([1] Bryant und Lawrence entfernen homologe Proteine nicht aus der Menge, auf der das Potential berechnet wird [382].)

Um den Einfluß von Speichereffekten einzuschränken, muß darauf geachtet werden, daß die Mengen, auf denen das Potential berechnet und getestet wird, disjunkt sind oder der Einfluß der jeweils getesteten Struktur aus dem Potential herausgerechnet wird, was durch den sogenannten *Jack-Knife*–Test erreicht werden kann.
Tabelle 5.1 zeigt die Selbsterkennungsrate für verschiedene Voronoipotentiale im Vergleich mit den in der Literatur für andere Potentiale zu findenden Werten. Zur Ableitung der Tabelle wurden die Voronoipotentiale auf den von den jeweiligen Autoren verwendeten Testmengen angewendet. Obwohl alle Voronoipotentiale einen geringen Detailgrad aufweisen (in [139] werden allein zwanzig Distanzintervalle verwendet), übertreffen sie bis auf ein Potential alle anderen Potentiale, was die Erkennung der nativen Struktur auf Rang $\mathcal{R}[\infty]$ angeht. Die einzige Ausnahme ist die Testmenge von Bryant und Lawrence [48], wo bei der Berechnung des Originalwertes homologe Proteine nicht aus der Trainingsmenge entfernt wurden. In umfangreichen Analysen [268, 269] verschiedener empirischer Potentiale zeigen Levitt *et al.* durch Selbsterkennungstests auf unterschiedlich generierten Testmengen, daß diese Potentiale erfolgreich zwischen verschiedenen Faltungen unterscheiden können doch starke Schwächen bei der Unterscheidung nativer Strukturen von nahezu–nativen Strukturen (*RMS*–Abweichung kleiner als 5.0Å) haben.
Diese Ergebnisse deuten daraufhin, daß eine Optimierung von Sequenzstrukturalignments allein anhand empirischer Paarinteraktionspotentiale schwierig, wenn

nicht sogar unmöglich ist. Anhand des im folgenden betrachteten Beispiels aus der JOY–Datenbank [306] struktureller Alignments wird dies sehr deutlich. Das Datenbankalignment für die Proteine 3phv und 2rspA weißt eine Identität von 26.3% auf, und die Superpositionierung gemäß des Alignments hat eine *RMS*–Abweichung von 2.66Å bei 93 alinierten Positionen. Aliniert man die Sequenz

```
2rspA  _000:-LAMTMEHK-DRPLVRVILTNTGSHPVKQRSVYITALLDSGADITIISEEDWPTDWPVME
3phv   _000:P-QITL---WQRPLVTIKIGGQLK----------EALLDTGADDTVLEEMSLPGRWKPKM
Allowed_000:..******..**********.............******************.......
2rspA  _000:~LAMTMEHK~DRPLVRVILTNTGSHPVKQRSVYITALLDSGADITIISEEDWPTDWPVME
3phv   _000:~PQITL--W~QRPLVTIKIG----------GQLKEALLDTGADDTVLEEMSLP--GRWKP

2rspA  _060:AA~~~~~~GI-PMRKS--RDMIELGVINRDGSLERPLLLFPAVAMVRGSILGRDCLQGLG
3phv   _060:IG~~~~~~GIGGFIKVRQYDQILIEICGHK-----AIGT-VLVGPTPVNIIGRNLLTQIG
Allowed_060:..................********.............********************
2rspA  _060:AA------GI~PMRKS~~RDMIELGVINRDGSLERPLLLFPAVAMVRGSILGRDCLQGLG
3phv   _060:KMIGGIGGFI~KVRQY~~-DQILIEIC------GHKAIGTVLVGPTPVNIIGRNLLTQIG

2rspA  _120:LRLTNL-
3phv   _120:CTLN--F
Allowed_120:****..
2rspA  _120:LRLTNL,
3phv   _120:CTLNF-,
```

Abbildung 5.4: Vergleich des JOY–Alignments (2rspA, 3phv) (oben) gegen das Sequenzstrukturalignment (unten).

von 3phv mit der RDP–Methode mit der Paarpotentialbewertung als maßgebliches Optimierungskriterium gegen die Struktur von 2rspA, so werden ebenfalls 93 Reste bei sogar 28.3% Identitäten aliniert. Die *RMS*–Abweichung beträgt jedoch 5.68Å. Vergleicht man die beiden Alignments positionell (siehe auch Abschnitt 8.3.1), so stimmen sie nur an 53.2% der Positionen (in Abbildung 5.4 mit * markiert) überein.
Ein typisches Paarpotential bewertet die native Struktur von 2rspA mit −38.3 und die Einbettung der 3phv–Sequenz in die 2rspA–Struktur gemäß des JOY–Alignment mit −27.0. Das schlechtere aber hinsichtlich der Paarpotentialbewertung optimierte Alignment liefert jedoch ein Paarpotentialenergie von −36.0 und erreicht damit nahezu die Bewertung der nativen Struktur. Dies zeigt, daß auch empirische Paarpotentiale genau wie Austauschmatrizen nicht allein als Optimierungskriterium für Sequenzstrukturalignments verwendet werden können.
Werden die in Kapitel 7 beschriebenen Parametrisierungen der Kostenfunktionen und Randbedingungen eingesetzt, so berechnet die RDP–Methode übrigens ein Alignment mit 3.3Å *RMS*–Abweichung, 27.3% Identitäten und einer Paarpotentialbewertung von −28.4.

Kapitel 6

Rekursive Dynamische Programmierung (RDP) als allgemeines Lösungskonzept für biologisch relevante Fragestellungen

6.1 Motivation

Die in dieser Arbeit entwickelte Methode der **R**ekursiven **D**ynamischen **P**rogrammierung (RDP) ist ein allgemeines algorithmisches Schema zur Lösung von Vergleichsproblemen, wie sie im Bereich der molekularen Biologie auftreten. Diese Lösungsmethode hat ihr Hauptanwendungsgebiet dort, wo es um die Lösung *NP–vollständiger* Probleme [108] – wie dem allgemeinen Sequenzstrukturalignment mit aminosäuretypabhängigen Wechselwirkungspotentialen – geht. Für diese Art von Problemen kann eine optimale Lösung nur garantiert werden, wenn man im schlechtesten Fall exponentielle Laufzeit in der Größe der Problembeschreibung in Kauf nehmen kann. Exponentielle Laufzeiten sind jedoch für die meisten biologisch relevanten Problemstellungen inakzeptabel, so daß man auf mit heuristischen Verfahren berechnete approximative Lösungen angewiesen ist. Bei biologisch relevanten Vergleichsproblemen ist jedoch die bezüglich der verwendeten Kostenfunktionen optimale Lösung in vielen Fällen nicht identisch mit der *biologisch richtigen* Lösung, so daß die Frage berechtigt erscheint, welchen Sinn die Berechnung der unter Kostenfunktionsaspekten *optimalen* Lösung macht. Das wesentliche Ziel der RDP–Methode ist, biologisch sinnvolle anstelle von Lösungen zu berechnen, die bezüglich ungenauer Kostenfunktionen oder Modelle optimal sind.
Bei typischen Problemstellungen in der Bioinformatik ist die Problemeigenschaft, daß optimale Gesamtlösungen effizient aus optimalen Lösungen für Teilprobleme zusammengesetzt werden können, ein wesentliches Kriterium dafür, daß optimale Lösungen effizient berechnet werden können. Diese Problemeigenschaft wird im wesentlichen durch die verwendete Kostenfunktion bestimmt. Wird der Kostenfunktionswert aus einem Vergleich von lokalen Eigenschaften der betrachteten Objekte errechnet, können in der Regel effiziente Lösungsverfahren gefunden werden. Dies ist zum Beispiel der Fall, wenn, wie beim Sequenzalignment, die Kostenfunktion aus numerischen Werten für die Ähnlichkeit beziehungsweise Austauschbarkeit zweier Aminosäuren und Bestrafungstermen für Insertionen beziehungsweise Deletionen besteht, und damit die Bewertung einer Zuordnung im Alignment unabhängig von anderen Zuordnungen im Alignment ist. Diese Lokalitätseigenschaft geht jedoch verloren, wenn – wie beim Sequenzstrukturalignment mit Paarinteraktionspotentialen – nicht nur lokale Eigenschaften sondern auch

globale Eigenschaften wie Wechselwirkungen zwischen in der Sequenz weit von einander entfernt liegenden Aminosäuren Bestandteil der Kostenfunktion sind.
Ist neben der Lokalität der Kostenfunktion auch noch eine Reihenfolge auf den betrachteten Objekten vorgegeben, wie dies bei Proteinsequenzen, die allgemeiner auch als lineare Folge von Zeichen über einem Alphabet angesehen werden können, natürlicherweise der Fall ist, so kann die Aufteilung von vorn in der Form von Präfixen erfolgen. Die Lösung des Gesamtproblems erfolgt dann über die schrittweise Berechnung von optimalen Lösungen für die Präfixe. Daher spricht man in diesem Fall auch von dem Prinzip der Präfixoptimalität. Präfixoptimalität bedeutet, daß Präfixe der optimalen Lösung des Gesamtproblems auch gleichzeitig optimale Lösungen der durch die Präfixe definierten Teilprobleme darstellen. Das Prinzip der Präfixoptimalität ist die grundlegende Voraussetzung zur Anwendbarkeit des beim Vergleich von Proteinsequenzen in der Regel eingesetzten Optimierungsverfahrens der dynamischen Programmierung (siehe dazu Abschnitt 4.1.2).

Lemma 6.1
Wenn beim Sequenzstrukturalignment (siehe Definition 4.5) aminosäuretypabhängige Wechselwirkungspotentiale Bestandteil der Kostenfunktion sind, dann ist für das zugehörige Optimierungsproblem das Kriterium der Präfixoptimalität verletzt.

Beweisskizze:
Bezeichne im folgenden $h(A_{|k}, B_{|l}) = h_{|k,l}$ ein Alignment der Präfixe $A_{|k}$ der Sequenz A und $B_{|l}$ der Struktur B. Sei $\mathcal{H} := \{h \mid h(A,B) : A \longrightarrow B\}$ und $\mathcal{H}(A_{|i}, B_{|j}) := \{h \mid h(A_{|i}, B_{|j}) : A_{|i} \longrightarrow B_{|j}\}$. Sei die Optimierung eine Maximierung von $Score(h)$ und seien $\mathcal{H}^{opt}$ und $\mathcal{H}^{opt}(A_{|i}, B_{|j})$ die entsprechenden Mengen der bezüglich der Bewertungsfunktion $Score$ optimalen Alignments.
Bezeichne α die Gaperöffnungskosten und β die Gapverlängerungskosten (vgl. Definition 4.1) und $Match_{|i,j}(i,j)$ die Kosten für eine Paarung von i mit der Position j ohne Einbeziehung von Kontakten zu Positionen $k > j$ in B.

Annahme: Das Sequenzstrukturalignmentproblem mit aminosäuretypabhängigem Wechselwirkungspotential erfüllt das Kriterium der Präfixoptimalität, das heißt: $\forall\ h \in \mathcal{H}^{opt} :\ h_{|i,j} \in \mathcal{H}^{opt}(A_{|i}, B_{|j})$

Sei das Paar (i,j) mit $i \in A$ und $j \in B$ nun so gewählt, daß

- $\forall\ h \in \mathcal{H}^{opt} : h(i) = j$ und
- (j,k) der einzige Kontakt von Position $j \in B$ ist, und
- $Match_{|i,j}(i,j) < 2 * (\alpha + \beta) \quad \wedge \quad Match(i,j) > 2 * (\alpha + \beta) \qquad \alpha, \beta \in \mathbb{R}$

Falls also der Kontakt (j,k) in die Bewertung einbezogen wird, alinieren alle optimalen Alignments die Positionen i und j. Falls er nicht einbezogen wird, werden

sowohl Position $i \in A$ als auch Position $j \in B$ von einem optimalen Alignment nicht aliniert und wenn nötig ein Wechselgap eingefügt.
$\Longrightarrow \forall\ h_{|i,j} \in \mathcal{H}^{opt}(A_{|i}, B_{|j}) :\ h_{|i,j}(i) = \emptyset\ \wedge\ h^{-1}_{|i,j}(j) = \emptyset$
$\Longrightarrow \exists\ h \in \mathcal{H} :\ h(i) \neq j\ \wedge\ \forall\ h' \in \mathcal{H} \backslash \{h\} \wedge h'(i) = j :\ Score\{h_{|i,j}\} > Score\{h'_{|i,j}\}$
$\Longrightarrow \exists\ h \in \mathcal{H} :\ h(i) \neq j\ \wedge\ \forall\ h' \in \mathcal{H}^{opt} :\ Score\{h_{|i,j}\} > Score\{h'_{|i,j}\}$
$\Longrightarrow \exists\ h' \in \mathcal{H}^{opt} :\ h'_{|i,j} \notin \mathcal{H}^{opt}(A_{|i}, B_{|j})$
$\Longrightarrow$ Widerspruch zur Annahme ■

Da die Anwendbarkeit der Methode der dynamischen Programmierung zur Berechnung optimaler Alignments auf der Eigenschaft der Präfixoptimalität des Problems beruht, folgt aus Lemma 6.1 direkt:

Satz 6.1
Wenn beim Sequenzstrukturalignment aminosäuretypabhängige Wechselwirkungspotentiale Bestandteil der Kostenfunktion sind, kann mit der Methode der dynamischen Programmierung die Optimalität der berechneten Abbildung der Sequenz auf die Struktur nicht garantiert werden.

Lathrop [193] hat – wie bereits erwähnt – bewiesen, daß das Sequenzstrukturalignmentproblem *NP–vollständig* ist, wenn aminosäuretypabhängige Wechselwirkungspotentiale Bestandteil der Kostenfunktion sind und Insertionen beziehungsweise Deletionen beliebiger Länge erlaubt sind. Damit kann unter der Annahme, daß $P \neq NP$ gilt, kein polynomieller Algorithmus (also auch nicht die Methode der dynamischen Programmierung) die optimale Lösung allgemeiner Sequenzstrukturalignmentprobleme garantieren.
Wird dennoch die Methode der dynamischen Programmierung zur Berechnung solcher Sequenzstrukturalignments eingesetzt, wie dies in vielen Ansätzen der Fall ist (siehe Abschnitt 4.2.3), so treten bei der Berechnung eines Sequenzstrukturalignments die zwei in Abbildung 6.1 gezeigten Probleme auf:
Das *Backward*–Problem und das *Forward*–Problem beschreiben unterschiedliche Sichten des gleichen Problems, da ein Kontakt (j', j), der an Position j ein *Backward*–Problem hervorruft, bereits an Position j' ein *Forward*–Problem zur Folge hatte. Dennoch unterscheiden sich die beiden Probleme aus berechnungstechnischer Sicht:

- Das *Backward*–Problem ist deshalb schwierig, weil es nicht mehr genügt, nur den Wert der optimalen Präfixalignments zu kennen. Vielmehr müssen auch die optimalen Präfixalignments selbst und zum gewissen Grad sogar suboptimale Alignments bei der Bewertung des Kontaktes (l, j) berücksichtigt werden, da ihre Bewertung von der zu Position l alinierten Aminosäure abhängt.

  In der Regel gibt es mehrere optimale Präfixalignments mit dem gleichen Kostenfunktionswert und auch zahlreiche suboptimale Alignments, deren

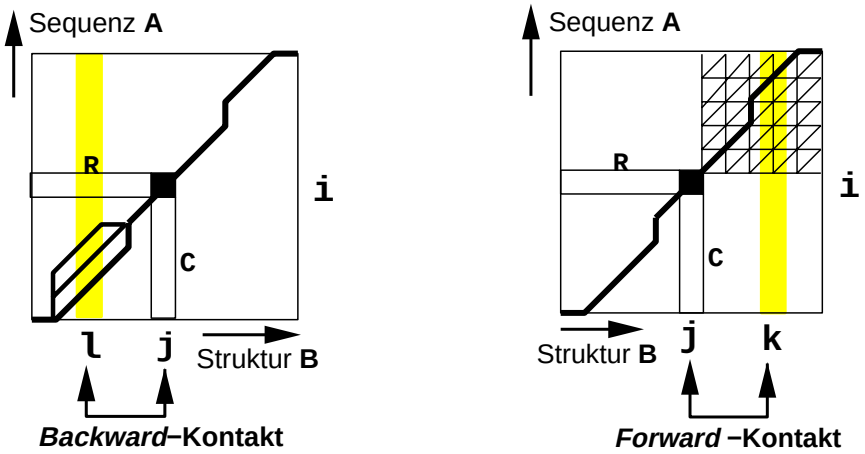


Abbildung 6.1: Das *Backward*–Problem (links, die Pfade zeigen optimale Präfixalignments) und das *Forward*–Problem (rechts).

Kostenfunktionswert den optimalen Wert nur um einen gewissen Wert unterschreitet. Im schlechtesten Fall sind exponentiell viele optimale und suboptimale Präfixalignments bei der Bewertung des Kontaktes (l, j) einzubeziehen. Falls durch Einbeziehung der aktuellen Position in die Bewertung eine Auftrennung der bisher gültige Präfixalignments erfolgt, ist die bisherige Menge der Präfixalignment entsprechend anzupassen.

Die Hoffnung ist jedoch, daß in praktischen Anwendungen die Anzahl der auszuwertenden Präfixalignments, wie in Abbildung 6.1 angedeutet, relativ klein ist, so daß sie eventuell entweder gespeichert oder rückberechnet werden könnten.

- Das *Forward*–Problem dagegen kann in keinem Fall aufgelöst werden. Zu dem Zeitpunkt, zu dem das Präfixalignment bis Position j berechnet wird, kann nicht nur eine begrenzte Anzahl von Pfaden in der dynamischen Programmiermatrix noch zur optimalen Gesamtlösung führen, sondern zur Vervollständigung des Gesamtalignments stehen noch alle prinzipiell möglichen und damit exponentiell viele Postfixalignmemts zur Auswahl. Der rechte Teil von Abbildung 6.1 verdeutlicht dieses Problem.

Ein zu Satz 6.1 analoges Resultat kann auch für das Strukturalignmentproblem gezeigt werden. Der Beweis kann ebenfalls indirekt geführt werden. Hier soll jedoch darauf verzichtet werden, da anschaulich klar ist, daß, egal wie man eine optimale Teilsuperposition bestimmt, diese nicht unbedingt zu einer optimalen globalen Superposition der Gesamtstruktur gehören muß. So superpositionieren zum Beispiel alle Helices gleicher Länge mehr oder weniger gleich gut. Welche Fehler ein inkrementeller Aufbau der Gesamtsuperposition aus Teillösungen zum Beispiel beim Vergleich zweier *TIM–barrel*–Strukturen zur Folge haben kann, ist anschaulich sehr klar.

In Abschnitt 4.2.1 sind bereits die in der Literatur zur Lösung des Sequenzstrukturalignmentproblems vorgestellten Ansätze und ihre jeweiligen Vor– und Nachteile diskutiert worden. Aufgrund des *NP*–Vollständigkeitsresultats ist klar, daß diese Verfahren entweder im schlechtesten Fall exponentielle Laufzeit haben oder aber Heuristiken sind, die die Optimalität der gefundenen Lösung nicht garantieren können.
Die verschiedenen Kostenfunktionbestandteile werden mit statistischen Methoden aus den jeweiligen Sequenz– und Strukturdatenbanken abgeleitet (siehe Abschnitte 5.1–5.2.4) und sind daher mit einer nicht zu unterschätzenden Fehlerquote behaftet. Bei der Lösung der meisten biologischen Vergleichsprobleme, wie zum Beispiel dem Sequenzstrukturalignmentproblem, steht daher weniger das Finden der exakten global optimalen Lösung im Vordergrund, sondern vielmehr kommt es darauf an, eine Methode zu entwickeln, die selbstadaptierend die teilweise sehr schwachen Signale in den biologischen Daten findet. Beim Entwurf der in dieser Arbeit beschriebenen RDP–Methode wurde daher besonderer Wert darauf gelegt, daß sich die Methode selbständig an die jeweils vorhandenen Informationen über das Problem und bereits festgelegte Teillösungen adaptiert. Für das Sequenzstrukturalignmentproblem bedeutet dies zum Beispiel, daß sich die Methode mit unterschiedlicher Priorität an die verschiedenen, zur Bewertung des Alignments beitragenden Bewertungsfunktionsterme wie Sequenzähnlichkeit (siehe Abschnitt 5.1) oder Paarpotentialbewertung (siehe Abschnitt 5.2.4) anpassen muß.

6.2 Die RDP–Methode

Aus den oben erörterten Gründen wird hier zur Lösung des Sequenzstrukturalignmentproblems ein *divide & conquer*–Algorithmus vorgeschlagen, der das Gesamtproblem rekursiv als eine Folge von disjunkten Teilproblemen löst und in verschiedenen Stadien des Verfahrens unterschiedliche Gewichtung auf die unterschiedlichen Kostenfunktionsbestandteile legt. Die RDP–Methode [340] verwendet zwei Phasen zur Lösung von Vergleichsproblemen: In der *top–down*–Phase erfolgt die rekursive Zerlegung des zu bearbeitenden Problems in kleinere Unterprobleme, so daß eine Hierarchie von Unterproblemen entsteht. Die anschließende *bottom–up*–Phase setzt die für diese Unterprobleme erhaltenen Teillösungen zu einer Gesamtlösung zusammen. Die Abbildung 6.2 zeigt schematisch den algorithmischen Ablauf der RDP–Methode.
In der *top–down*–Phase des Algorithmus werden für ein zu bearbeitendes Problem P rekursiv auf jeder Hierarchieebene folgende Berechnungschritte durchgeführt:

Bestimmung signifikanter Teillösungen: In der Prozedur *Oracle* werden für ein Problem P die zuverlässig zuzuordnenden Bereiche der zu vergleichenden Objekte bestimmt. Zur Bestimmung dieser Bereiche werden in der Regel verschiedene Methoden und Kostenfunktionsmodelle eingesetzt, so

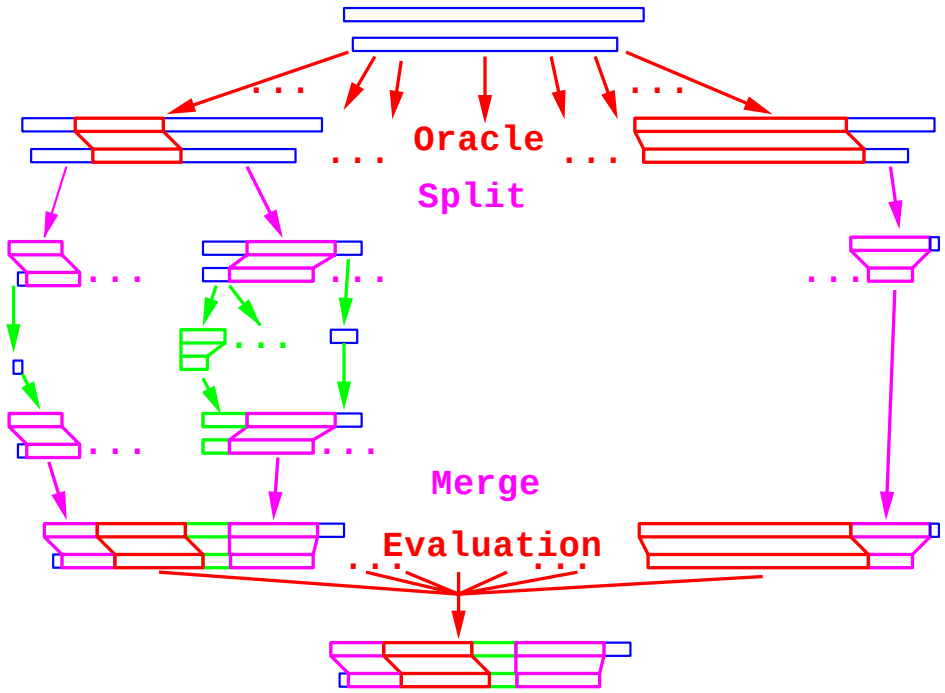


Abbildung 6.2: Schema der RDP–Methode am Beispiel eines paarweisen Sequenzvergleiches: Der *Oracle*–Schritt bestimmt die zuverlässig zuzuordnenden Bereiche. Der *Split*–Schritt zerlegt das Teilproblem in disjunkte Unterprobleme anhand der k Teillösungen aus dem *Oracle*–Schritt. Der *Merge*–Schritt kombiniert die Lösungen für ein Teilproblem mit den Lösungen der Unterprobleme. Im *Evaluation*–Schritt werden die so entstandenen n Lösungen neubewertet und die besten m an die darüberliegende Hierarchiestufe weitergereicht.

daß der *Oracle*–Schritt eine Menge von optimalen oder auch suboptimalen Zuordnungsmöglichkeiten generiert. Die Realisierung dieser Funktion ist abhängig von der Art des zu bearbeiteten Problems und wird exemplarisch für das Sequenzstrukturalignmentproblem in Kapitel 7 detaillierter beschrieben.

Aufteilung in Teilprobleme auf Basis der Teillösungen : Im *Split*–Schritt wird das aktuell bearbeitete Teilproblem entsprechend einer Teillösung oder einer Menge von Teillösungen in noch zu bearbeitende Unterprobleme zerlegt. Die Teilobjekte, die bereits aufeinander abgebildet werden konnten und die aktuelle Zerlegung definieren, werden aus den noch zu bearbeitenden Teilobjekten entfernt, so daß eine Auftrennung in kleinere Teilprobleme erfolgt.

Abbildung 6.3 beschreibt die Vorgehensweise des *Split*–Schrittes. Die Funktion

Split $(\mathcal{T}, P, A, pq)$
  $\mathcal{P}_A \longleftarrow$ *decompose* (P, A)
  for $SP \in \mathcal{P}_A$ & $SP \neq \emptyset$ **do**
    $pq \longleftarrow$ *queue_insert* (pq, SP)
    $\mathcal{T} \longleftarrow$ *tree_insert* $(\mathcal{T}, SP)$
  return $[\mathcal{T}, pq]$

Abbildung 6.3: Ablaufschema der Funktion *Split* (Pseudocode in Anlehnung an [68]). P und SP sind Probleminstanzen und $\mathcal{P}_A$ ist die Menge der Unterprobleme von P, die durch das Alignment A definiert werden.

decompose führt eine Aufteilung des Teilproblems gemäß des jeweiligen Alignments in Unterprobleme SP durch. Sofern die so entstandenen Unterprobleme SP nicht leer sind, werden sie in den Lösungsbaum $\mathcal{T}$ und in die Liste der noch zu bearbeitenden Teilprobleme pq eingefügt. Noch nicht bearbeitete Teilprobleme werden im folgenden auch als *offene* Teilprobleme bezeichnet.
Der dynamisch auf– und abgebaute Lösungsbaum $\mathcal{T}$ dient zur Verwaltung sowohl der offenen Teilprobleme SP und der zugehörigen Teilobjekte als auch der bereits bestimmten Teillösungen. Eine genauere Beschreibung dieser für die RDP–Methode zentralen Datenstruktur erfolgt in Abschnitt 7.1.5. Abbildung 6.4 zeigt

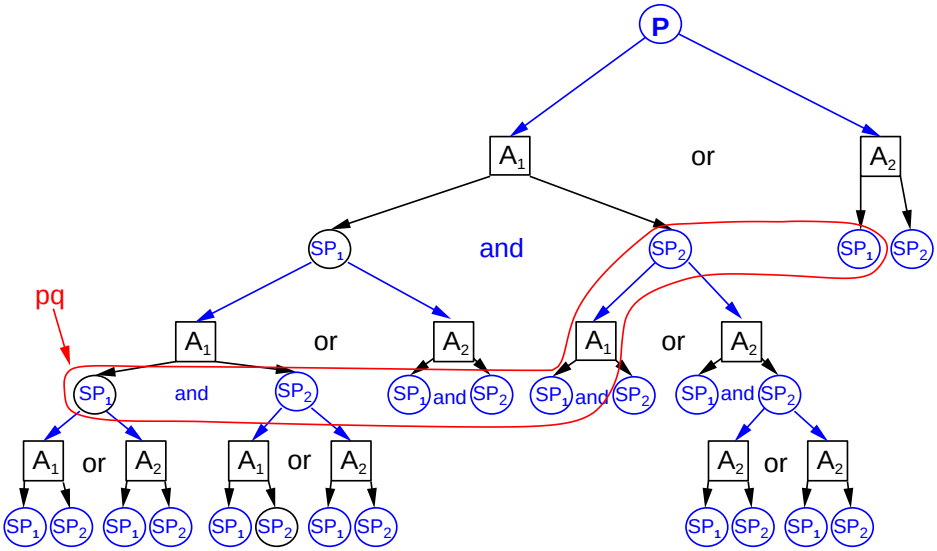


Abbildung 6.4: Der Lösungsbaum $\mathcal{T}$ und die Verwaltung der offenen Teilprobleme in der Prioritätsliste pq. Teillösungen sind als rechteckige Knoten A_i und Unterprobleme als runde Knoten P beziehungsweise SP_i dargestellt.

diesen Baum $\mathcal{T}$ und den Inhalt der Prioritätsliste pq in einem beliebigen Zustand

des Algorithmus. Neue Knoten werden in den Baum $\mathcal{T}$ sowohl bei der Berechnung von Teillösungen im *Oracle*–Schritt (in Abbildung 6.4 als rechteckige Knoten mit A_i gekennzeichnet) als auch durch die Aufteilung in Unterprobleme im *Split*–Schritt (in Abbildung 6.4 als runde Knoten mit SP_j gekennzeichnet) eingefügt. Die Löschung von Knoten aus dem Baum erfolgt in der *bottom–up*–Phase, die später beschrieben werden wird.
Die Liste *pq* der offenen Teilprobleme ist in Form einer Prioritätsliste organisiert, die nach vom Typ der bearbeiteten Problemstellung abhängigen Kriterien sortiert ist. Ein einfaches Kriterium kann zum Beispiel die Größe der Teilprobleme sein. In der Regel ist eine sinnvollere Abarbeitungsreihenfolge abhängig vom Typ der bearbeiteten Problemstellung. Für das Sequenzstrukturalignmentproblem werden verschiedene Strategien in Abschnitt 7.5 diskutiert.
Die Zerlegung wird solange rekursiv fortgesetzt, bis entweder die zu bearbeitenden Teilprobleme leer sind oder im jeweiligen *Oracle*–Schritt zum Beispiel keine Teillösung zur Aufteilung in weitere Unterprobleme mehr gefunden wird, deren positionelle Bewertung signifikant höher ist als die einer globalen Lösung des gesamten Restproblems. Der Bezug der Bewertung auf die Anzahl der alinierten Positionen ist sinnvoll, da insbesondere Paarinteraktionsbewertungen von der Anzahl der betrachteten Positionen abhängig sind [366].
Um die Wahrscheinlichkeit von Fehlern zu verringern, die durch falsche Entscheidungen auf höheren Ebenen der Rekursion entstehen, wo noch wenig über den Rest der in Berechnung befindlichen Gesamtlösung bekannt ist, werden zur Lösung eines jeden Teilproblems nicht nur unterschiedliche *Oracle*–Prozeduren eingesetzt, sondern auch neben den jeweils optimalen Lösungen suboptimale Lösungen verfolgt. Obwohl dadurch der abgesuchte Lösungsraum vergrößert wird und damit der Berechnungsaufwand zunimmt, beschränkt die RDP–Methode die Suche effizient auf den Teil des Lösungsraums, der sich durch biologisch relevante Ähnlichkeiten und damit signifikante Teillösungen auszeichnet, wie die Ergebnisse in Kapitel 8 für das Sequenzstrukturalignment zeigen.
Nachdem alle Probleme eines Teilbaumes in der *top–down*–Phase gelöst sind, ergibt sich die Gesamtlösung für das Teilproblem, das durch diesen Teilbaum repräsentiert wird, aus der Kombination der in dem Teilbaum gespeicherten Teillösungen. In der *bottom–up*–Phase des Algorithmus erfolgt das Zusammensetzen der Gesamtlösung durch die Funktion *Finish* (siehe Abbildung 6.5). Die Funktion *Finish* durchläuft den Baum $\mathcal{T}$ in *depth first search*–Reihenfolge und führt für jedes Problem P die folgenden Prozeduren aus:

Zusammenfassung von Teillösungen: In dem *Merge*–Schritt (siehe Abbildung 6.6) wird die Gesamtlösung des jeweiligen Teilproblems durch Kombination der lokalen Teillösung, die die Aufteilung in Unterprobleme definiert hat, mit den für die zugehörigen Unterprobleme ermittelten Lösungen bestimmt. Die Unterprobleme werden danach samt der zu ihnen gehörigen Unterlösungen aus dem Baum $\mathcal{T}$ entfernt.

Finish$(\mathcal{T}, P)$
  for each $A \in P.\mathcal{A}$ **do**
    for each $SP \in \mathcal{P}_A$ **do**
      $[\mathcal{T}, SP] \longleftarrow$ ***Finish***$(\mathcal{T}, SP)$
  $[\mathcal{T}, P.\mathcal{A}] \longleftarrow$ ***Merge***$(\mathcal{T}, P)$
  $P.\mathcal{A} \longleftarrow$ ***Evaluation***$(\mathcal{T}, P, P.\mathcal{A})$
  return $[\mathcal{T}, P]$

Abbildung 6.5: Die Funktion *Finish* (in Pseudocode) durchläuft in *depth first search*–Reihenfolge den Baum $\mathcal{T}$ und ruft die Funktionen *Merge* (siehe Abbildung 6.6) und *Evaluation* für jeden Knoten auf. $SP.\mathcal{A}$ ist die Menge der bisher berechneten Alignments für das Teilproblem SP.

Auswahl weiterverwendeter Teillösungen: Im *Evaluation*–Schritt erfolgt zunächst eine Auswahl der so entstandenen Lösungen nach Kriterien, die als zusätzliche Randbedingungen in das Verfahren einfließen. Beim Sequenzstrukturalignment sind dies zum Beispiel Kriterien, die die Konsistenz der durch das Alignment implizierten Modellstruktur und somit die Verwendbarkeit des Alignments für eine spätere ähnlichkeitsbasierte Modellierung sicherstellen. Lösungen, die diesen Kriterien nicht genügen, werden verworfen. Von den restlichen Lösungen werden nach einer Neubewertung die mit dem besten Bewertungsfunktionswert an die darüberliegende Hierarchiestufe weitergereicht.

Merge $(\mathcal{T}, P)$
  $\mathcal{A} \longleftarrow P.A$
  for each $A \in P.\mathcal{A}$ **do**
    for each $SP \in \mathcal{SP}_A$ **do**
      for $SA \in SP.\mathcal{A}$ **do**
        $\mathcal{A} \longleftarrow$ *join* $(\mathcal{A}, A, SA)$
      $\mathcal{T} \longleftarrow$ *tree_remove*$(\mathcal{T}, SP)$
  return $[\mathcal{T}, \mathcal{A}]$

Abbildung 6.6: In der Funktion *Merge* werden lokale Lösungen A aus P.$\mathcal{A}$ und die Lösungen SA für die Unterprobleme SP zu Lösungen für das Teilproblem P zusammengesetzt.

Die Notwendigkeit einer Neubewertung der Teillösungen in der *bottom–up*–Phase erklärt sich am Beispiel des Sequenzstrukturalignments daher, daß zum Zeitpunkt der Zusammensetzung der Teillösungen mehr Information über die Gesamtlösung

und damit über die Verträglichkeit der Teillösung mit einem in Konstruktion befindlichen Strukturmodell existiert, als dies zum Zeitpunkt der Bestimmung der Teillösungen der Fall war.

Die Auswahl der an die darüberliegende Hierarchiestufe weitergereichten Teillösungen kann über zwei Kriterien beeinflußt werden. Zum einem kann die Anzahl über einen Parameter des Verfahrens eingestellt werden. Zum anderen kann die Einschränkung anhand eines Signifikanzgrenzwertes erfolgen, dessen Definition von der Art des bearbeiteten Problems abhängt. Der *Evaluation*–Schritt ist wie die Prozedur *Oracle* stark abhängig von der Art des bearbeiteten Problems und wird für das Sequenzstrukturalignment in Abschnitt 7 detaillierter beschrieben.

Abweichend von der normalen Abarbeitungsreihenfolge in der *bottom–up*–Phase des RDP–Algorithmus wird der Lösungsbaum $\mathcal{T}$ nicht nur dynamisch expandiert, sondern kann in vielen Problemstellungen in Teilen wieder kontrahiert werden, bevor der gesamte Baum aufgebaut ist. Dies ist zu dem Zeitpunkt möglich, zu dem alle zur endgültigen Bewertung der Teillösungen benötigten Informationen vorhanden sind. Dies kann zum Beispiel aufgrund von Lokalitätseigenschaften der eingesetzten Kostenfunktionen der Fall sein, bevor alle Teilprobleme bearbeitet sind. Beim Sequenzstrukturalignment zum Beispiel ist dies der Fall, wenn alle externen, das heißt nicht in dem Teilproblem befindlichen Kontaktpartner bekannt sind.

Während die Zerlegungsphase endet, wenn die Unterprobleme leer sind oder keine signifikanten lokalen Teillösungen mehr gefunden werden, endet die Kombinationsphase dann, wenn eine oder mehrere Lösungen für das Gesamtproblem vorliegen.

```
RDP_main (P)
    T ⟵ new_tree (P)
    pq ⟵ init_queue (P)
    RDP (T, pq)
    for each A ∈ P.A do
        output (A)

RDP (T, pq)
    if pq ≠ ∅ do
        SP ⟵ select (pq)
        for each Oracle do
            SP.A ⟵ Oracle (SP, T)
        for each A ∈ SP.A do
            [T, pq] ⟵ Split (T, SP, A, pq)
        for each SP ∈ T & ready (SP) do
            [T, SP] ⟵ Finish (T, SP)
        RDP (T, pq)
    else
        [T, P] ⟵ Finish (T, P)
```

Abbildung 6.7: Ablaufschema des RDP–Algorithmus (Pseudocode in Anlehnung an [68]). P und SP bezeichnen Probleminstanzen. $SP.\mathcal{A}$ ist die Menge der für SP berechneten Alignments.

Abbildung 6.7 gibt eine zusammenfassende Beschreibung der Ablaufstruktur des RDP–Algorithmus. In der Prozedur *new_tree* wird der Baum zur Verwaltung der noch nicht abgeschlossenen Teilprobleme angelegt und initialisiert. Zu Beginn der Prozedur besteht dieser Baum aus der Wurzel und das der Wurzel assoziierte Problem ist das aktuell bearbeitete Problem. Die Prozedur *init_queue* erzeugt die Liste, in der die offenen Teilprobleme verwaltet werden, und fügt das Gesamtproblem als erstes Teilproblem ein.
Das Attribut *ready* gibt für ein Teilproblem SP an, ob zum einen alle Unterprobleme dieses Teilproblems bereits abgeschlossen sind und zum anderen alle zur Bewertung der der Teillösungen notwendigen Informationen vorliegen. Bei der Bearbeitung von Sequenzstrukturalignmentproblemen, bei denen aminosäuretypabhängige Paarinteraktionspotentiale Bestandteil der Kostenfunktion sind, erfordert dieses Attribut zum Beispiel, daß zu diesem Zeitpunkt alle Interaktionspartner festgelegt sind.
Die Ausprägung der Funktion *Oracle* ist grundlegend von der bearbeiteten Anwendung abhängig und ist der Kern der RDP–Methode. Daher ist ihr auch der Abschnitt 7.2 in dieser Arbeit gewidmet. Hier sei zunächst nur gesagt, daß diese Funktion partielle Zuordnungen für die das Teilproblem definierenden Objekte liefert, die eine Zerlegung der aktuellen Teilproblems in disjunkte Unterprobleme erlaubt.

Kapitel 7

Sequenzstrukturalignment mit RDP

In Kapitel 6 wurde der allgemeine Aufbau der RDP–Methode beschrieben und gezeigt, warum Standardverfahren, wie die dynamische Programmierung, für viele biologisch relevante Vergleichsprobleme nicht ohne weiteres erfolgreich einsetzbar sind. In diesem Kapitel wird die konkrete Realisierung der RDP–Methode für ein spezielles biologisches Vergleichsproblem, das Sequenzstrukturalignment, beschrieben. Die Anwendung der RDP–Methode auf das Sequenzstrukturalignmentproblem ist aus folgenden Gründen interessant:

- Die ähnlichkeitsbasierte Proteinstrukturvorhersage ist der gegenwärtig erfolgversprechendste Ansatz zur theoretischen Proteinstrukturvorhersage (siehe Abschnitt 3.3).
- Das Sequenzstrukturalignment ist die zentrale Komponente in der ähnlichkeitsbasierten Proteinstrukturvorhersage (siehe Abschnitte 4.1 und 4.2).
- Das Problem ist derzeit mit herkömmlichen Methoden nur unzureichend gelöst. Das betrifft sowohl die verwendeten Bewertungssysteme (siehe Kapitel 5) als auch die Algorithmen zur Lösung des zugehörigen Optimierungsproblems, das in vielen Ausprägungen *NP–vollständig* ist (siehe Abschnitt 4.2).

Die formale Definition 4.5 des Sequenzstrukturalignmentproblems sieht auf den ersten Blick sehr einfach aus: Gesucht wird ein partieller injektiver Homomorphismus zwischen einer Sequenz und einer Struktur, der eine bestimmte Kostenfunktion optimiert.
Bis heute ist jedoch keine Kostenfunktion bekannt, deren Optimierung in der realen Anwendung die beste Lösung im Sinne der Verwendbarkeit des aus dem Alignment resultierenden Modells liefert. Daher ist es beim heutigen Wissensstand ratsam, bei der Berechnung von Sequenzstrukturalignments alle verfügbaren, biologischen und biochemischen Informationen einzubeziehen. Diese Informationen sind heterogen und unterscheiden sich von Anwendungsfall zu Anwendungsfall. Die folgende Liste gibt eine Übersicht der in der RDP–Methode zur Berechnung von Sequenzstrukturalignments eingesetzten Kostenfunktionsbestandteile und Randbedingungen:

- Zu den in der RDP–Methode zur Bewertung und Berechnung von Lösungen eingesetzten *Bewertungssystemen* (siehe Kapitel 5) gehören:
  - Austauschmatrizen (siehe Abschnitt 5.1), die die Sequenzähnlichkeit bewerten,

  - Hydrophobizitätspotentiale (siehe Abschnitt 5.2.3.1), die eine Präferenz für Aminosäuretypen kodieren, mit dem Lösungsmittel zu interagieren oder vergraben zu sein,
  - Kontaktkapazitätspotentiale (siehe Abschnitt 5.2.3.2), die für einen Aminosäuretypen die Wahrscheinlichkeit angeben, in einer bestimmten Kontaktumgebung aufzutreten,
  - Zwei– und Mehrkörperpotentiale (siehe Abschnitt 5.2.4), die Wechselwirkungen bestimmter Aminosäurereste innerhalb der dreidimensionalen Struktur bewerten.

- *Diskrete Zusatzinformationen* werden hauptsächlich in Form von Randbedingungen zur Auswahl von Teillösungen verwendet, zum Beispiel:

  - Schwefelbrücken, die durch die beteiligten Cysteinreste definiert werden und damit die räumliche Lage dieser Reste zueinander im Strukturmodell vorschreiben,
  - geladene / polare Aminosäuren im hydrophoben Kern des Proteins, die nach Möglichkeit im Strukturmodell einen geeigneten Wechselwirkungspartner haben sollten,
  - aus Mutationsexperimenten abgeleitete Wechselwirkungen bestimmter Aminosäuren.

- Wissen über die *Funktion* des Proteins kann als Randbedingung für die Auswahl von Lösungen aber auch zur Ableitung von Teillösungen in der RDP–Methode verwendet werden:

  - Bekannte Bindungsstellen für Liganden können sowohl eine bestimmte Konformation des Strukturmodells vorschreiben, aber auch in Sequenz und Struktur konserviert sein, so daß eine direkte Zuordnung der beteiligten Reste erfolgen kann,
  - Aminosäurereste, deren Lösungsmittelzugänglichkeit aus Experimenten bekannt ist oder aus ihrer Funktion abgeleitet werden kann, sollten auch im aus dem Alignment resultierenden Strukturmodell zugänglich sein.

- *Strukturelle* beziehungsweise *geometrische Randbedingungen* werden zur Auswahl zulässiger Teillösungen eingesetzt, zum Beispiel

  - werden Insertionen nur an zugänglichen Stellen erlaubt, da ansonsten das resultierende Strukturmodell nicht realisierbar ist,
  - muß bei Deletionen darauf geachtet werden, daß die entstehenden Lücken im Strukturmodell durch Schleifen schließbar sind,
  - müssen Abstandsbedingungen aus *Crosslinker*– beziehungsweise Mutationsexperimenten eingehalten werden.

Die RDP–Methode hebt sich von anderen Sequenzstrukturalignmentmethoden unter anderem dadurch ab, daß sie

- diese verschiedenen Kriterien direkt in die Berechnung einbezieht,
- alle Hinweise auf mögliche Ähnlichkeiten zwischen Sequenz– und Strukturteilen nutzt und so
- neben der lösungsspezifischen Signifikanzbewertung eine unabhängige Bestätigung der Hinweise durch unterschiedliche Kriterien erhält und damit
- adaptiv die jeweils relevanten Signale detektieren kann.

Die verschiedenen Informationen gehen dabei entweder in die bei der Generierung und Bewertung von Teillösungen verwendeten Kostenfunktionen oder als Randbedingungen bei der Auswahl signifikanter und zulässiger Teillösungen ein.

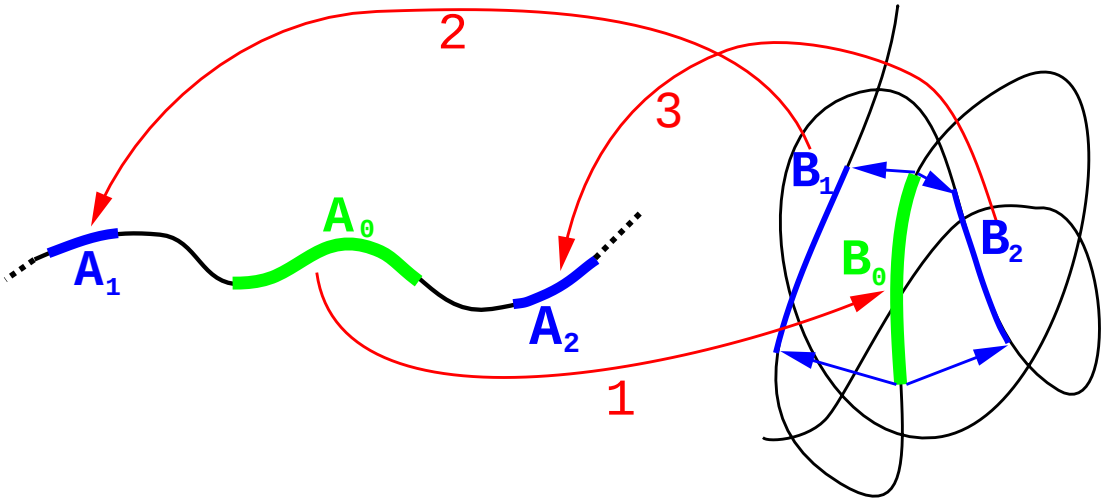


Abbildung 7.1: Die RDP–Methode in Anwendung auf das Sequenzstrukturalignment: Gesucht ist eine Abbildung der Sequenz A auf die Struktur B (die Schritte 1 bis 3 werden im Text erläutert).

Abbildung 7.1 verdeutlicht noch einmal den groben Ablauf der RDP–Methode am Anwendungsbeispiel des Sequenzstrukturalignments. Die Abbildung zeigt, wie die RDP–Methode ein vollständiges Alignment einer Sequenz A mit einer Struktur B berechnet:

1. Suche die beste lokale Zuordnung (Alignment). Die Funktion *Oracle* bestimmt eine oder mehrere gemäß der heterogenen Kriterien signifikante Teillösungen $A_0 \longrightarrow B_0$.

2. Zerlege das Gesamtproblem in Teilprobleme (*Split*), durch Herausschneiden der bereits gefundenen Zuordnung oder durch Identifikation der mit bereits zugeordneten Bereichen kontaktierenden Teile der Struktur B_i.

3. Suche für die so erzeugten Teilprobleme wiederum nach guten lokalen Zuordnungen (*Oracle*), zum Beispiel nach Teilen der Sequenz A_i, die gut gegen das durch ein Paarpotential ausgebildete Interaktionsprofil alinieren.

4. Bilde die so gefundenen Teile $A_i \longrightarrow B_i$ ab und fahre rekursiv mit 2. fort.

5. Stoppe, falls kein signifikantes *Teil*alignment oder andere signifikante Teilzuordnungen mehr identifizierbar sind.

Die folgenden Abschnitte beschreiben, wie die skizzierte RDP–Methode die verschiedenen Kriterien zur Berechnung von Sequenzstrukturalignments nutzt, und demonstrieren damit die Flexibilität und Effizienz der RDP–Methode.
Dieses Kapitel gliedert sich wie folgt: In Abschnitt 7.1 werden die verwendeten Datenstrukturen eingeführt. Die wesentlichen Schritte der *top–down*–Phase der RDP–Methode werden in den Abschnitten 7.2 bis 7.4 vorgestellt: Abschnitt 7.2 beschreibt, wie in den verschiedenen Ausprägungen der Funktion *Oracle* die verschiedenen Kostenfunktionen und Randbedingungen zur Generierung von Teillösungen für Teilprobleme verwendet werden. In Abschnitt 7.3 werden verschiedene Filter vorgestellt, die die von den verschiedenen Orakeln generierten Teillösungen auf die zulässigen und signifikanten Teillösungen einschränken und identische Teillösungen direkt eliminieren. Außerdem wird in Abschnitt 7.3 eine allgemeine Methode vorgestellt, die bereits in der *top–down*–Phase mehrere miteinander verträgliche Teillösungen zu einer Teillösung zusammenfaßt und so den weiteren Berechnungsaufwand reduziert. In Abschnitt 7.4 wird die Aufteilung von Problemen in Unterprobleme (*Split*) anhand von bisher gefundenen Teillösungen beschrieben, und in Abschnitt 7.5 wird die Abarbeitungsreihenfolge der offenen Teilprobleme diskutiert.
Der Abschnitt 7.6 beschäftigt sich genauer mit der *bottom–up*–Phase der RDP–Methode: Unterabschnitt 7.6.1 beschreibt die Zusammensetzung von Teillösungen (*Merge*), das heißt wie aus den Lösungen von Unterproblemen Gesamtlösungen berechnet werden. Die Bewertung und Auswahl von Teillösungen (*Evaluation*), die bei der Zusammensetzung der Gesamtlösung an die jeweils darüberliegende Hierarchiestufe weitergereicht werden, erfolgt in Abschnitt 7.6.2.
In Abschnitt 7.7 wird das Problem der Gewichtung der verschiedenen Kostenfunktionsbestandteile adressiert und werden die für die Ergebnisse in Kapitel 8 verwendeten Parametereinstellungen bestimmt.

7.1 Datenstrukturen

Die RDP–Methode verwendet zur Berechnung von Sequenzstrukturalignments im wesentlichen folgende Datenstrukturen:

- Der Lösungsbaum $\mathcal{T}$ (siehe Abschnitt 6.2) ist die zentrale Datenstruktur der RDP–Methode. In ihm werden die Teilprobleme und bereits gefundene Teillösungen verwaltet.

- Die Prioritätsschlange dient der signifikanzbasierten Steuerung der Abarbeitungsreihenfolge der Teilprobleme.

Daneben gibt es Datenstrukturen zur Beschreibung der Sequenz und der Struktur beziehungsweise Faltung einschließlich der zu ihnen bekannten Zusatzinformationen. Teilprobleme werden durch Abbildungen kodiert, die die zum Teilprobleme gehörigen Teilsequenzen und –strukturen in der Gesamtsequenz beziehungsweise –struktur identifizieren (siehe Abschnitt 7.1.3).

7.1.1 Informationen zur Sequenz

Abbildung 7.2 zeigt die von der RDP–Methode zur Speicherung der zur Sequenz gehörigen Informationen. Die Beschreibung der Aminosäuresequenz erfolgt in ei-

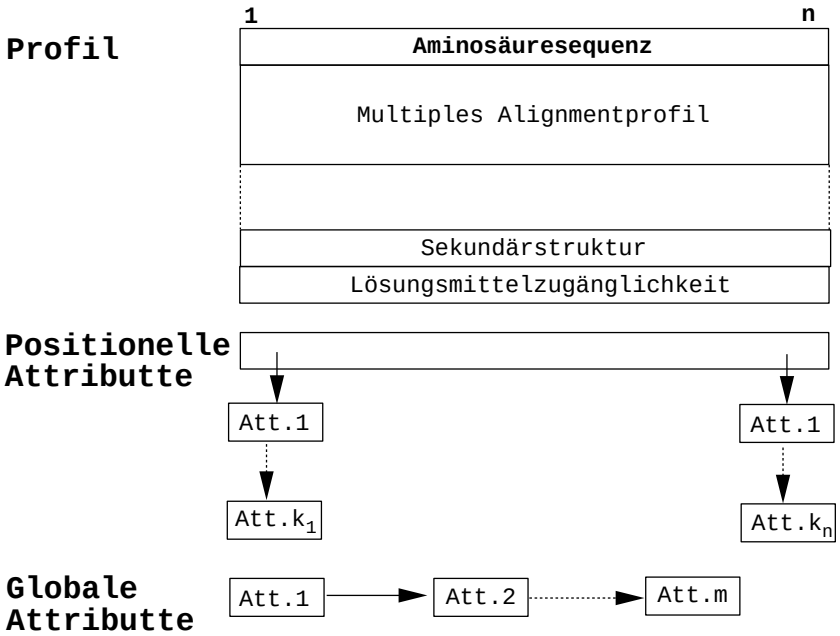


Abbildung 7.2: Datenstruktur der zur Sequenz gehörigen Informationen.

nem sogenannten Profil aus beliebig vielen Vektoren der Länge der Sequenz. Im einfachsten Falle also aus einem Vektor, der die Aminosäuresequenz speichert. In weiteren Vektoren können dann sowohl einfache positionelle Annotationen, wie zum Beispiel vorhergesagte Sekundärstrukturzuordnungen im Dreizustandsmodell (Helix, Strang, Schleife) oder Lösungsmittelzugänglichkeiten, aber auch ein vorberechnetes multiples Alignment von mit der Sequenz verwandten Sequenzen abgelegt werden, so daß eine einfache Erweiterung der RDP–Methode auf multiple Sequenzen möglich ist.
In den Listen der *positionellen Attribute* werden komplexere Eigenschaften von Sequenzpositionen gespeichert. Zum Beispiel werden Aminosäurereste, die an der Ausbildung bekannter aktiver Stellen oder Schwefelbrücken beteiligt sind,

mit einen Bezeichner für die Art der aktiven Stelle beziehungsweise dem an der Brücke beteiligten Partner annotiert. Die Zuordnung dieser Informationen zu den Aminosäureresten ermöglicht in den verschiedenen Auswahlfunktionen der RDP–Methode den effizienten Test, ob bei der Bearbeitung eines Teilproblems oder bei der Auswahl von Teillösungen neben der Kostenfunktion besondere Regeln oder zusätzliche Randbedingungen zu berücksichtigen sind.
Kann eine Eigenschaft der Sequenz nicht bestimmten Aminosäureresten zugeordnet werden, wird diese Information in die Liste der *globalen Attribute* eingetragen, die bei den verschiedenen Auswahlmechanismen der RDP–Methode ausgewertet wird.

7.1.2 Informationen zur Struktur

Zur Repräsentation einer Proteinstruktur mit Länge m wird die Datenstruktur aus Abbildung 7.2 (Sekundärstruktur und Lösungsmittelzugänglichkeit werden der `DSSP`–Datenbank [173] entnommen) erweitert um

- eine Matrix (Dimension $3 \times m$) zur Speicherung der Koordinaten der C_α–Atome (zum Beispiel benötigt zur Auswertung von Randbedingungen) und
- m Adjazenzlisten, in denen die Distanz– oder Voronoikontaktrelationen gepeichert werden.

Die Adjazenzlisten der Distanz– und Voronoikontaktrelationen unterscheiden sich nur durch ihre Berechnung und die jeweils gespeicherten Attribute der Kontaktrelationen. Gegenwärtig werden nur Paarkontaktrelationen verwendet, eine Verwendung von Mehrkörperwechselwirkungen ist jedoch in der RDP–Methode möglich, indem anstelle der Kanten des Kontaktrelationsgraphen (siehe Definition 5.6) *Hyper*kanten verwendet werden.
Die Berechnung von Distanzkontaktrelationen erfolgt beim Einlesen einer neuen Struktur, da die zu berücksichtigenden Kontakte vom jeweiligen empirischen Potential abhängen. Zum Beispiel hat ein Potential, das nur Kontakte mit einem maximalen Distanzabstand von 10Å bewertet, wesentlich weniger Kontakte als eines mit 20Å Maximalabstand, wie er zum Beispiel bei den von Sippl vorgeschlagenen Potentialen [319] verwendet wird. Abbildung 5.3 zeigt die Zunahme der Anzahl der Kontakte mit steigender Kontaktdistanz.
Die Voronoikontaktrelationen haben dagegen den großen Vorteil, daß sie vorberechnet und in einer Datenbank abgelegt werden können, da die Voronoikontaktrelation auf echter räumlicher Nachbarschaft beruht und damit distanzunabhängig ist. Der räumliche Abstand der Kontaktpartner und die Kontaktfläche sind mögliche zusätzliche Attribute einer Voronoikontaktrelation. Bei der Verwendung von Voronoikontaktrelationen wird außerdem die Lösungsmittelzugänglichkeit aus der Voronoizerlegung berechnet.

Die Bewertung von Kontaktrelationen unter Verwendung der den beteiligten Strukturpositionen bereits zugeordneten Aminosäurereste ist zentraler Bestandteil sowohl der Schritte der RDP–Methode, die Teillösungen generieren, als auch der Schritte, die Lösungen bewerten und auswählen. Zur Effizienzsteigerung werden daher die Bestandteile der Bewertungsfunktion, die nur von Attributen des Kontaktes und nicht von den daran beteiligten Aminosäuretypen abhängen, wie zum Beispiel vom Abstand der beteiligten Positionen in der Aminosäurekette (Sequenzabstand) und im Raum (euklidischer Abstand), vor der eigentlichen Optimierung berechnet und als zusätzliche Attribute der Kontaktrelation zugeordnet.

7.1.3 Darstellung von Teilproblemen

Ein (Teil–)Problem ist beim Sequenzstrukturalignment eindeutig durch die zu ihm gehörigen Sequenz– und Strukturbereiche definiert. Am Anfang der rekursiven Zerlegung sind dies die vollständige Sequenz und die vollständige Struktur.

Definition 7.1 (Teilproblem)
Ein Teilproblem SP *des zu berechnenden Sequenzstrukturalignments von Sequenz* $A =< a_1, \ldots, a_n >$ *auf Struktur* $B =< b_1, \ldots, b_m >$ *wird definiert durch die Teile* $A|_{SP} =< a_{\pi_A^{SP}(1)}, \ldots, a_{\pi_A^{SP}(k)} >$ *der Sequenz* A *der Länge* k *und den Teil* $B|_{SP} =< b_{\pi_B^{SP}(1)}, \ldots, b_{\pi_B^{SP}(l)} >$ *der Struktur* B *der Länge* l*, die Bestandteil des Teilproblems* SP *sind.*

Die Abbildungen π_A^{SP} und π_B^{SP} beschreiben die Zuordnung von Positionen der Teilsequenz $A|_{SP}$ beziehungsweise Teilstruktur $B|_{SP}$ zu Positionen der Gesamtsequenz A beziehungsweise Gesamtstruktur B.
Der Lösungsbaum der RDP–Methode enthält zur Laufzeit zahlreiche Teilprobleme. Daher werden Teilprobleme SP speicherplatzeffizient beschrieben durch

- die jeweiligen Anfangs– und Endpositionen in der Sequenz und der Struktur, falls keine Inversionen von Sequenz– und Strukturbereichen erlaubt sind und Teilprobleme aus zusammenhängenden Segmenten bestehen, oder sonst durch
- zwei Indexvektoren $\pi_{\{A,B\}}^{SP}$ der Länge der Teilsequenz/–struktur, die die Abbildung der Teile auf die Gesamtsequenz/–struktur beinhalten.

Über diesen Indizierungsmechanismus wird auf alle Datenstrukturen zugegriffen, die die Gesamtsequenz und die Gesamtstruktur beschreiben. Unzusammenhängende Struktur– oder Sequenzsegmente können zum Beispiel durch das Ausblenden von Schleifenbereichen oder von Bereichen niedriger Sequenzkomplexität (zum Beispiel in der Verbindung zwischen zwei Domänen [37]) entstehen. Inversionen von Strukturbereichen treten zum Beispiel dann auf, wenn es erlaubt ist, die Verbindungstopologie der Peptidkette dynamisch zu modifizieren. Derartige

Veränderungen der Peptidkette haben sich jedoch bisher als nicht sinnvoll erwiesen.
Im Regelfall kann ein Teilproblem im Lösungsbaum daher durch vier Zahlen vollständig beschrieben werden. Zur Vereinfachung der Schreibweise wird die Abbildung $\pi^{SP}_{\{A,B\}}$ im folgenden häufig weggelassen, und die Aminosäurereste werden stattdessen durch ihre Position in der Gesamtsequenz oder –struktur adressiert.

7.1.4 Darstellung von Teillösungen

Definition 7.2 (Teillösung)
Eine Teillösung $f|_{SP}$ *ist eine auf das Teilproblem SP eingeschränkte Lösung des Sequenzstrukturalignments (siehe Definition 4.5), also ein partieller injektiver Homomorphismus* $f|_{SP} : A|_{SP} \longrightarrow B|_{SP}$.
Wenn im folgenden das bearbeitete Teilproblem SP aus dem Kontext klar hervorgeht, werden $f' = f|_{SP}$, $A' = A|_{SP}$ *und* $B' = B|_{SP}$ *als Abkürzungen verwendet*

Da während der RDP–Methode zahlreiche Teillösungen generiert werden und teilweise erst in der *bottom–up*–Phase des Algorithmus verworfen werden können, muß die zur Speicherung von Teillösungen verwendete Datenstruktur nicht nur eine effiziente Bewertung unterstützen, sondern auch speicherplatzeffizient sein.

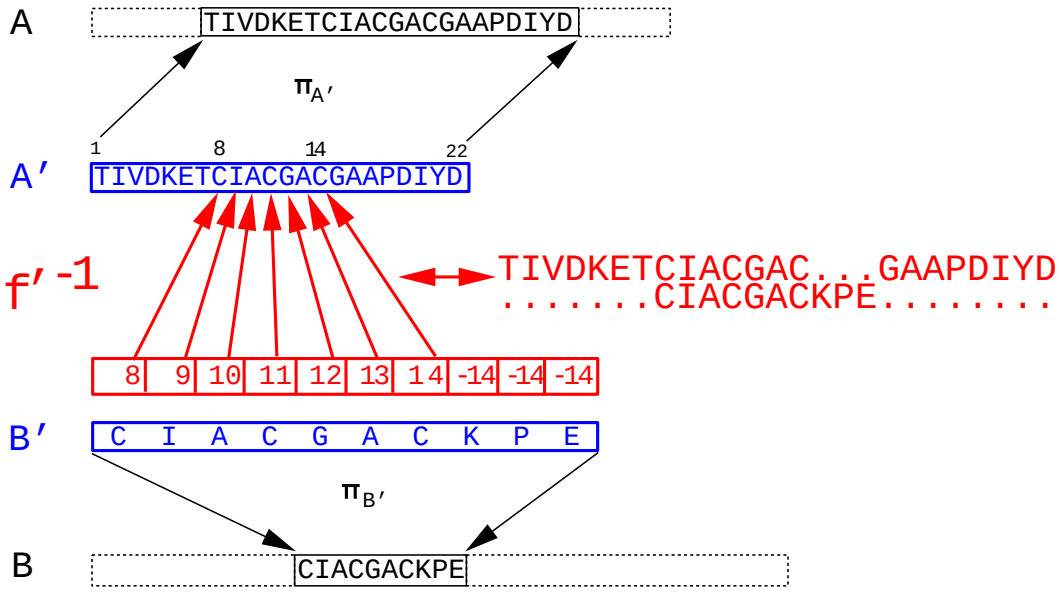


Abbildung 7.3: Speicherung von Teilproblemen und Teillösungen: A' und B' definieren ein Teilproblem für Sequenz A und Struktur B. Der Vektor f'^{-1} kodiert eine Teillösung für das Teilproblem.

Alle Teillösung beziehungsweise Alignments werden daher, wie in Abbildung 7.3 gezeigt, durch einen Vektor von Indices kodiert, dessen Länge der Teilstruktur des Teilproblems entspricht.

Falls $f^{-1}|_{SP}(i) \neq \emptyset$, gibt der Eintrag i an, welcher Aminosäurerest der Sequenz der Strukturposition i zugeordnet wird. Ansonsten kodiert ein negativer Eintrag, daß die Position durch die Teillösung $f|_{SP}$ unbelegt bleibt. Sein Wert gibt dabei zusätzlich die Sequenzposition an, nach der die Deletion (bezogen auf die Struktur) beginnt, so daß zum Beispiel bei Wechselgaps die Konvertierung zwischen der externen, textuellen Darstellung eines Alignments und seiner internen Darstellung eindeutig ist.
Diese Darstellung ermöglicht nicht nur die Speicherung von Teillösungen, die aus zusammenhängenden Bereichen bestehen, sondern erlaubt auch Insertionen und Deletionen, wie sie zum Beispiel Ergebnis lokaler Alignments sein können. Zudem ermöglicht diese Kodierung die Darstellung der von Orakeln generierten Lösungen und von Teillösungen, die durch Zusammenfassen in der *bottom–up*–Phase entstehen, in einheitlicher Form. Bei der Zusammenfassung von Teillösungen werden zuvor negative Einträge durch die Zuordnungen ersetzt, die durch die Lösungen der Unterprobleme bestimmt sind.
Eine der elementaren Funktionen der RDP–Methode ist die Bewertung von Teillösungen anhand der bereits mit Aminosäureresten der Sequenz belegten Kontaktrelationen. Um diese effizient durchführen zu können, wird intern die Umkehrabbildung f^{-1} statt f verwendet, da so bei der Bewertung einer Kontaktrelation mit dem Index der daran beteiligten Strukturpositionen über f^{-1} direkt auf den Aminosäuretyp und andere Informationen aus der Sequenz zugegriffen werden kann. Bei Verwendung von f müßte dagegen im schlechtesten Fall der ganze Vektor durchlaufen werden, um festzustellen, ob und welche Sequenzposition einer Strukturposition von der Teillösung zugeordnet wird.

7.1.5 Der Lösungsbaum $\mathcal{T}$

Der Lösungsbaum $\mathcal{T}$ (siehe auch Abbildungen 6.4 und 7.6) dient nicht nur zur Speicherung von Teilproblemen und Teillösungen, sondern durch seine Topologie wird auch die Verknüpfung zwischen den verschiedenen Teillösungen kodiert. $\mathcal{T}$ ist ein dynamischer *Und–Oder*–Baum und wird über seine Generierungsfunktionen definiert:

Definition 7.3 (Lösungsbaum)
Der Lösungsbaum $\mathcal{T} = (V_\mathcal{T}, E_\mathcal{T})$ ist ein Baum mit Knotenmenge $V_\mathcal{T} = V_\wedge \cup V_\vee$:

- *Ein Knoten $v \in V_\vee$ ist ein Tupel $(SP_v^\vee, TA_v^\vee)$, wobei $SP_v^\vee$ die Beschreibung des Teilproblems und $TA_v^\vee$ die Menge der Lösungen für das Teilproblem ist. Die Wurzel (P, A) des Lösungsbaums $\mathcal{T}$ ist in $V_\vee$.*
- *Ein Knoten $v \in V_\wedge$ ist ein Tupel $(PA_v^\wedge, TA_v^\wedge)$, wobei $PA_v^\wedge$ eine* partielle *Lösung und $TA_v^\wedge$ eine Menge von Lösungen für ein Teilproblem ist.*

Der Lösungsbaum $\mathcal{T}$ wird durch folgende Funktionen generiert:

- *Die Funktion* $g_\wedge : V_\vee \longrightarrow V_\wedge^m$ *generiert aus einem Knoten* $u \in V_\vee$ *eine Menge mit* m *Knoten aus* $V_\wedge$.
- *Die Funktion* $g_\vee : V_\wedge \longrightarrow V_\vee^n$ *generiert aus einem Knoten* $u \in V_\wedge$ *eine Menge mit* n *Knoten aus* $V_\vee$.

Die Kantenmenge $E_\mathcal{T}$ *ist definiert als* $E_\mathcal{T} = E^{down}_{\wedge\to\vee} \cup E^{down}_{\vee\to\wedge} \cup E^{up}_{\wedge\to\vee} \cup E^{up}_{\vee\to\wedge}$ *mit:*

- $E^{down}_{\vee\to\wedge} = \{(u,v) | u \in V_\vee \ \wedge \ v \in V_\wedge \ \wedge \ v \in g_\wedge(u)\}$
- $E^{down}_{\wedge\to\vee} = \{(u,v) | u \in V_\wedge \ \wedge \ v \in V_\vee \ \wedge \ v \in g_\vee(u)\}$
- $E^{up}_{\wedge\to\vee} = \{(u,v) | u \in V_\wedge \ \wedge \ v \in V_\vee \ \wedge \ u \in g_\wedge(v)\}$
- $E^{up}_{\vee\to\wedge} = \{(u,v) | u \in V_\vee \ \wedge \ v \in V_\wedge \ \wedge \ u \in g_\vee(v)\}$

Abbildung 7.4 zeigt eine Skizze des Lösungsbaums $\mathcal{T}$ und die darauf definierten Grundoperationen.

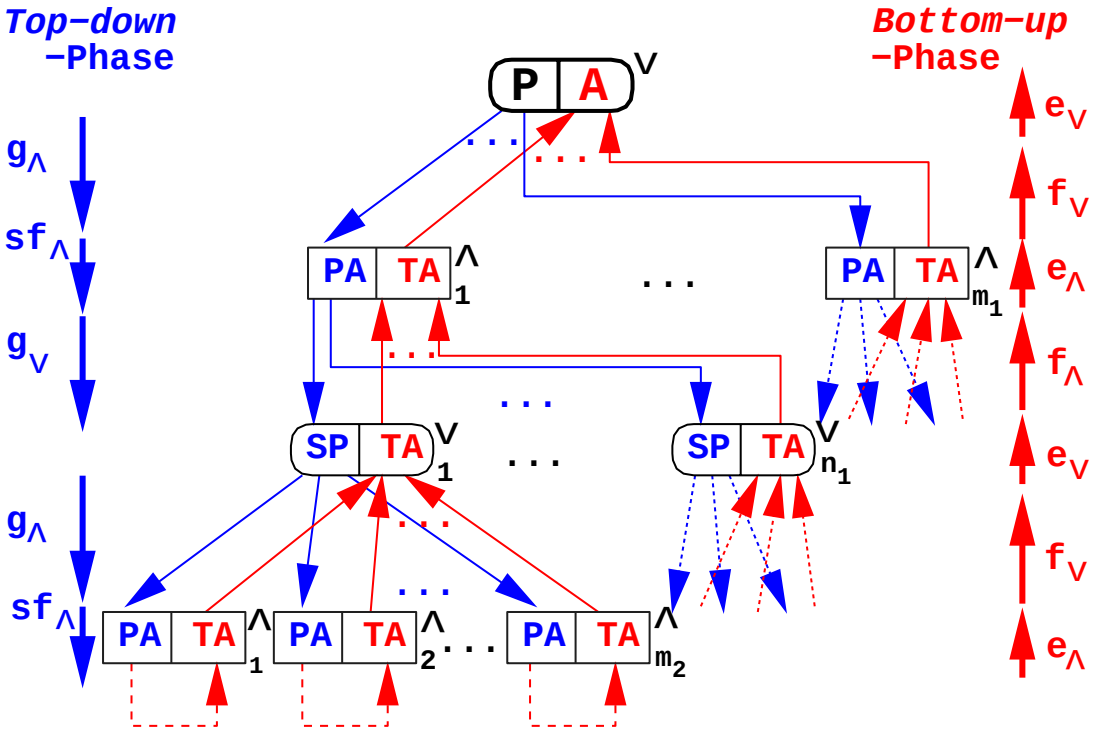


Abbildung 7.4: Die Datenstruktur des Lösungbaumes $\mathcal{T}$: Die *Und*–Knoten aus $V_\wedge$ sind als rechteckige Knoten und die *Oder*–Knoten aus $V_\vee$ sind als ovale Knoten dargestellt. Die Funktionen der *top–down*– und der *bottom–up*–Phase sind im Text beschrieben.

Beim Sequenzstrukturalignment mit der RDP–Methode repräsentieren die Knoten $v \in V_\mathcal{T}$ die Teilprobleme und deren Lösungen. In der *top–down*–Phase enthalten die Knoten $v \in V_\vee$ die Beschreibung von Teilproblemen $SP_v^\vee$ (siehe Abschnitt 7.1.3) und noch keine Lösung für das Teilproblem, das heißt $TA_v^\vee = \emptyset$. Die

Knoten $v \in V_\wedge$ enthalten in der *top–down*–Phase in $PA_v^\wedge$ eine partielle Zuordnung von Resten der Sequenz zu Positionen der Struktur (siehe Abschnitt 7.1.4) und die Menge der Lösungen $TA_v^\wedge$ ist ebenfalls leer.
Wird die Funktion $g_\vee$ auf einen Knoten $v \in V_\wedge$ angewendet, dann erzeugt sie neue Knoten $s = (SP_s^\vee, \emptyset)$ und fügt sie als Knoten aus $V_\vee$ in den Lösungsbaum $\mathcal{T}$ ein. Die neuen Unterprobleme $SP_s^\vee$ entstehen dabei durch Zerlegung des Teilproblems $SP_p^\vee$ mit $(p, v) \in E_{\vee\to\wedge}^{down}$ anhand der partiellen Zuordnung $PA_v^\wedge$. Ein Knoten $v \in V_\wedge$ ist terminaler Knoten des Lösungsbaums, wenn $g_\vee(v) = \emptyset$, also durch die Zerlegung nur leere Unterprobleme oder Unterprobleme entstehen, für die keine signifikante Lösung mehr gefunden werden kann. Da die Funktion $g_\vee$ zu einer Aufteilung in Unterprobleme führt, wird sie auch als *Split*–Funktion bezeichnet. Eine ausführliche Beschreibung erfolgt in Abschnitt 7.4.
Die zweite den Lösungsbaum $\mathcal{T}$ generierende Funktion $g_\wedge$ wird auf Knoten $u \in V_\vee$ angewendet, die noch zu lösende Teilprobleme beschreiben. $g_\wedge$ generiert für das Teilproblem $SP_u^\vee$ eine Menge von verschiedenen, partiellen Lösungen $PA_s^\wedge$, erzeugt für jede dieser $PA_s^\wedge$ einen Knoten $s = (PA_s^\wedge, \emptyset)$ und fügt diesen als Sohn von u in den Lösungsbaum $\mathcal{T}$ ein. Wird für ein Teilproblem $SP_u^\vee$ keine partielle Zuordnung mehr gefunden, das heißt $g_\wedge(u) = \emptyset$, dann ist u ein Blatt des Baumes und die Expansion des Lösungsbaums bricht an diesem Knoten ab. Diese partiellen Lösungen werden in der RDP–Methode auf unterschiedliche Weise unter Nutzung verschiedener Kriterien berechnet. Daher handelt es sich bei der Funktion $g_\wedge$ auch eigentlich um eine Menge von Funktionen, die im weiteren auch als *Orakel* bezeichnet werden. Die verschiedenen beim Sequenzstrukturalignment mit RDP verwendeten Orakel beschreibt der Abschnitt 7.2.
Ein Knoten $v \in V_\vee$ ist terminaler Knoten des Lösungsbaums, wenn die Orakel für das Teilproblem $SP_v^\vee$ keine zulässigen signifikanten partiellen Zuordnungen mehr liefern ($g_\wedge(v) = \emptyset$). Durch die an eine partielle Zuordnung gestellten Zulässigkeitskriterien (siehe Abschnitt 7.6.2) wird zum Beispiel sichergestellt, daß ein berechnetes Sequenzstrukturalignment in ein gültiges Strukturmodell mit durchgehender Peptidkette überführt werden kann.
Um von vorneherein Teilbäume des Lösungbaumes abzuschneiden, die durch nicht signifikante oder aufgrund von Randbedingungen nicht zulässige oder identische, beziehungsweise sehr ähnliche partielle Lösungen entstehen, werden die Knoten, die diesen Lösungen entsprechen, von der Funktion $sf_\wedge$, die auf Knoten aus $V^\wedge$ arbeitet, direkt wieder aus dem Lösungsbaum gelöscht. Beim Sequenzstrukturalignment ist die Signifikanz einer partiellen Lösung nur schwer getrennt von dem sie erzeugenden Orakel zu bewerten, daher wird die Signifikanz partieller Lösungen nicht nur in Abschnitt 7.3, sondern auch zusammen mit der Beschreibung des jeweiligen Orakels in Abschnitt 7.2 diskutiert. Die Erkennung und Behandlung identischer oder sehr ähnlicher partieller Lösungen wird in Abschnitt 7.3 dargestellt.
Damit sind die Operationen beschrieben, die in der *top–down*–Phase den Lösungsbaum aufbauen. Neben Funktionen $g_\wedge$ und $g_\vee$ gibt es Funktionen, die in der

bottom–up–Phase der RDP–Methode den Lösungsbaum $\mathcal{T}$ wieder zusammenfalten, abbauen und dabei die eine (oder eventuell auch mehrere) für das bearbeitete Problem von RDP berechnete Gesamtlösung aus der Menge der in $\mathcal{T}$ durch die *top–down*–Phase generierten partiellen Lösungen zusammensetzen.
Die *bottom–up*–Phase beginnt damit, daß für alle terminalen Knoten $v \in V_\wedge$ des Lösungsbaums $\mathcal{T}$ die partiellen Lösungen $PA_v^\wedge$ zu vollständigen Lösungen $TA_v^\wedge$ des Teilproblems umgeformt werden. Für terminale Knoten $v \in V_\vee$ wird die leere Lösung (zum Beispiel das leere Alignment) in die Menge der Lösungen $TA_v^\vee$ eingetragen. Danach wird der Lösungsbaum $\mathcal{T}$ durch die folgenden Funktionen schrittweise zusammengefaltet und abgebaut:

- Die Funktion $f_\vee : V_\vee \times V_\wedge^m \longrightarrow V_\vee$ faltet m Knoten aus $V_\wedge$ zu einer Lösung $l \in TA_v^\vee$ am Knoten $v \in V_\vee$ zusammen und löscht danach die zugehörigen m Knoten aus $V_\wedge$ und dem Lösungsbaum $\mathcal{T}$.
- Die Funktion $f_\wedge : V_\wedge \times V_\vee^n \longrightarrow V_\wedge$ faltet n Knoten aus $V_\vee$ zu einer Lösung $l \in TA_u^\wedge$ am Knoten $u \in V_\wedge$ zusammen und löscht danach die zugehörigen n Knoten aus $V_\vee$ und dem Lösungsbaum $\mathcal{T}$.

Die Funktion $f_\vee$ ist genau dann für einen Knoten $v \in V_\vee$ anwendbar, wenn für alle m Söhne $u \in V_\wedge$ von v die Lösungen $TA_u^\wedge$ berechnet worden sind. In der gegenwärtigen Implementierung der RDP–Methode wird in der Funktion $f_\vee$ die Menge der Lösungen $TA_v^\vee$ für Teilprobleme $SP_v^\vee$ durch Vereinigung der $TA_u^\wedge$ über alle Söhne u von v berechnet. Zur Zeit wird also in $f_\vee$ keine Evaluierung und Selektion von Lösungen durchgeführt.
Die Funktion $f_\wedge$ ist genau dann für einen Knoten $v \in V_\wedge$ anwendbar, wenn für alle n Söhne $u \in V_\vee$ von v die Einträge $TA_u^\vee$ berechnet worden sind. Der Hauptschritt der Funktion $f_\wedge$ besteht in der Berechnung der Menge der Lösungen $TA_v^\wedge$ für das Teilproblem, dessen Sohn v im Lösungsbaum $\mathcal{T}$ ist, durch Kombination der partiellen Lösung $PA_v^\wedge$ mit den Teillösungen $TA_u^\vee$ der Söhne u von Knoten v im Lösungsbaum $\mathcal{T}$. Abschnitt 7.6.1 beschreibt, wie diese kombinatorische Zusammenfassung in der Funktion *Merge* für das Sequenzstrukturalignment erfolgt.
Damit durch die rekursive Vereinigung aller Lösungsalternativen in $f_\vee$ und kombinatorische Zusammensetzung von Lösungen in $f_\wedge$ die Anzahl der Lösungen, die zur Wurzel des Lösungsbaums weiter gereicht werden, nicht exponentiell zunimmt, erfolgt durch die Funktionen $e_\wedge : V_\wedge \longrightarrow V_\wedge$ und $e_\vee : V_\vee \longrightarrow V_\vee$ eine Auswahl dieser Lösungen. In der gegenwärtigen Implementierung findet eine Auswahl von Lösungen nur an Knoten $v \in V_\vee$ durch die Funktion $e_\vee$ statt und $e_\wedge$ ist die Identitätsfunktion. Abschnitt 7.6.2 beschreibt, wie für einen Knoten $v \in V_\vee$ die Menge der Lösungen $TA_v^\vee$ beim Sequenzstrukturalignment mit der RDP–Methode eingschränkt wird.
Damit sind auch die in der *bottom–up*–Phase der RDP–Methode auf den Lösungsbaum $\mathcal{T}$ angewendeten Funktionen beschrieben. Zusammenfassend enthält ein Teilbaum, dessen Wurzel ein Knoten $v^\wedge = (PA_v^\wedge, TA_v^\wedge)$ aus $V_\wedge$ ist, zu jedem Zeit-

punkt alle bisher berechneten partiellen und ausgewählten vollständigen Lösungen für Teilprobleme, die sich aus der partiellen Lösung $PA_v^\wedge$ selbst oder aus partiellen Lösungen für Unterprobleme ergeben. Außerdem enthält der Teilbaum die Knoten, die die zugehörigen Teilprobleme und eventuell noch offenen Teilprobleme repräsentieren.
Alle Teillösungen, die zu einem bestimmten Teilproblem geführt haben, liegen auf dem Weg vom Teilproblem zur Wurzel des Baumes. Alle Teillösungen, die mit einer Teillösung zu einer größeren Lösung kombiniert werden können, befinden sich in Teilbäumen, die an Knoten aus $V_\wedge$, die partielle Lösungen repräsentieren, von dem Weg zur Wurzel abzweigen.
Über die Kanten des Lösungsbaums kann so zwischen den verschiedenen Teilproblemen und bereits gefundenen Teillösungen einfach navigiert werden. Dies ist insbesondere von Bedeutung, wenn bei Orakeln mit Mehrkörperpotentialen in der Bewertungsfunktion die für andere Teilprobleme bereits gefundenen Teillösungen bei der Lösung des aktuell bearbeiteten Problems einfließen müssen.

7.2 Generierung von Teillösungen

In der *top–down*–Phase der RDP–Methode wird die Funktion $g_\wedge$ auf die Knoten $v \in V_\vee$ des Lösungsbaums $\mathcal{T}$ angewendet (siehe Abschnitt 7.1.5).
Zur Abkürzung der Schreibweise wird im folgenden bei der Beschreibung der *top–down*–Phase ein Teilproblem $SP_v^\vee$ mit dem Knoten $v \in V_\vee$ gleichgesetzt und kurz als $SP \in \mathcal{T}$ bezeichnet. Ebenso wird die partielle Lösung $PA_v^\wedge$ mit dem Knoten $v \in V_\wedge$ gleichgesetzt und kurz als Teillösung $f' \in \mathcal{T}$ bezeichnet.
Die Berechnung partieller Lösungen f' wird in der RDP–Methode durch die Funktion *Oracle* (siehe Kapitel 6) implementiert. Die Funktion *Oracle* steht dabei zusammenfassend für verschiedene, im folgenden beschriebene Methoden oder auch Orakel, deren Gemeinsamkeit darin besteht, daß sie für ein Teilproblem SP partielle Lösungen f' liefern.

7.2.1 Orakel und Sequenzinformation

In der RDP–Methode werden zwei Strategien verwendet, um auf der Sequenzebene konservierte Bereiche zwischen Sequenz und Struktur zu identifizieren:

- Suche nach Bereichen hoher lokaler Sequenzähnlichkeit möglichst ohne Gaps.
- Globales oder *free–shift*–Alignment mit der Möglichkeit von Gaps mit anschließender Auswahl zuverlässiger Bereiche.

Jede dieser Vorgehensweisen wird für eine Auswahl der Aminosäureaustauschmatrizen angewendet, die in der Literatur vorgeschlagen und in Abschnitt 5.1 vorgestellt wurden. Es werden unterschiedliche Austauschmatrizen gleichzeitig verwendet, da zum einen für unterschiedliche Matrizen, insbesondere in Fällen

niedriger Sequenzähnlichkeit, auch unterschiedliche optimale Alignments berechnet werden, und zum anderen eine qualitative Rangordnung der Austauschmatrizen nur schwer unabhängig vom Anwendungsfall durchgeführt werden kann. Wenn die verwendeten Matrizen nicht anderweitig angegeben werden, werden die Matrizen `dayhoff` [73], `gonnet` [123], `blosum22` und `blosum50` [142] verwendet, die sich auch in den verschiedenen in der Literatur durchgeführten Vergleichsexperimenten als die besten Matrizen herausgestellt haben (siehe Abschnitt 5.1).

7.2.1.1 Lokale Sequenzalignments als Orakel

Die Suche nach Bereichen hoher lokaler Sequenzähnlichkeit zwischen Sequenz und Struktur kann auf zwei Arten erfolgen:

- Anwendung von `BLAST` [10].
- Berechnung von lokalen Alignments mit hohen Gapkosten.

Das Programm `BLAST` sucht nach zusammenhängenden Bereichen hoher Sequenzähnlichkeit und kann somit als ein Orakel in der RDP–Methode eingesetzt werden. Da der Einsatz von `BLAST` den Aufruf eines externen Programms und zusätzlichen Aufwand zur Konvertierung von Datenformaten bedeutet, werden, wenn `BLAST` als Orakel verwendet wird, einmal zu Beginn alle `BLAST`–Alignments zwischen der Sequenz A und der Struktur B berechnet und zusammen mit dem von `BLAST` berechneten Signifikanzindex (*p–value* [175]) in einer Liste gespeichert. In der zentralen und laufzeitkritischen Funktion der RDP–Methode kann somit ein `BLAST`–Orakelaufruf für ein Teilproblem durch Auswahl geeigneter `BLAST`–Alignments aus der Liste erfolgen.
Alternativ wird in der RDP–Methode für ein Teilproblem SP das lokale Alignment (siehe Abschnitt 4.1.2) zur Suche nach Bereichen hoher lokaler Sequenzähnlichkeit eingesetzt. Dynamische Programmierung mit affinen Gapkosten ist vom Berechnungsaufwand zwar quadratisch in der Eingabelänge, doch der mit dem Aufruf eines externen Programms verbundene Aufwand wiegt diesen höheren Berechnungsaufwand auf. Außerdem bedeutet der Einsatz der dynamischen Programmierung auch einen wesentlichen Zugewinn an Flexibilität.
Die für die verschiedenen sequenzbasierten Orakel verwendete Kostenfunktion

$$\phi(f', A', B') \;=\; \phi^S(f', A', B') \;+\; GAP(f', A', B')$$

besteht aus einem Anteil $\phi^S(f', A', B')$, der die Zuordnung auf Sequenzebene konservierter Bereiche belohnt, und aus den affinen Gapkosten $GAP(f', A', B')$. Die Zuordnung konservierter Bereiche wird wie im normalen Sequenzalignment durch den Kostenfunktionsterm

$$\phi^S(f', A', B') \;=\; \sum_{f'(i)\neq\emptyset} sim(a_i, b_{f'(i)})$$

bewertet. Die Funktion *sim* bewertet die Ähnlichkeit zweier Aminosäuren entsprechend der für das Orakel verwendeten Austauschmatrix. Die affinen Gapkosten $GAP(f', A', B')$ werden im folgenden analog zu Definition 4.1 durch den Term

$$GAP(f', A', B') \;=\; \alpha\,|G_{f'}(A', B')| \;+\; \beta\,|D_{f'}(A', B') \cup I_{f'}(A', B')|$$

beschrieben, wobei die Mengen $G_{f'}(A', B')$, $D_{f'}(A', B')$ und $I_{f'}(A', B')$ wie in Definition 3.3 definiert sind. Insertionen und Deletionen werden mit dem Faktor β und die Eröffnung von Gaps mit dem Faktor α bestraft.

Typischerweise werden in den Orakeln der RDP–Methode sowohl für die Eröffnung von Gaps als auch für ihre Verlängerung hohe Kosten angesetzt. Damit wird gewährleistet, daß die von den Orakeln generierten Teillösungen keine Gaps enthalten. Beim Sequenzstrukturalignment ist dies gleichbedeutend damit, daß in dem aus der Teillösung abgeleiteten Teilstrukturmodell keine Schleifen neu zu modellieren sind.

Es können aber nicht nur zusammenhängende Bereiche, also Bereiche ohne jede Deletion oder Insertion, gesucht werden, sondern über eine Variation der Gapkosten auch lokale Alignments, in denen Insertionen und Deletionen in gewissem Umfang erlaubt sind. Die Möglichkeit, Insertionen und Deletionen einzufügen, kann zur Folge haben, daß ein oder alle Orakel nur unzulässige Teillösungen (siehe Abschnitt 7.6.2) erzeugen. Diesem Problem wird von der RDP–Methode begegnet, indem die Gapkosten von der RDP–Methode schrittweise erhöht werden, bis zumindest einige der von den Orakeln generierten Lösungen zulässig sind oder klar ist, daß für das gegenwärtige Teilproblem keine zulässige Lösung existiert, die gleichzeitig noch den Signifikanzkriterien (siehe Abschnitt 7.3.1) genügt.

Ein weiterer Vorteil der Verwendung der dynamischen Programmierung zum Auffinden lokaler Teillösungen auf Sequenzebene besteht darin, daß bei gleicher asymptotischer Laufzeit nicht nur eine optimale, sondern alle optimalen und sogar suboptimale Lösungen für ein Teilproblem berechnet werden. Eine suboptimale Lösung ist eine Lösung, deren *Score* vom *Score* der optimalen Lösungen höchstens um einen vorgegebenen Prozentsatz abweicht.

Lokale Alignments identifizieren zwar immer die am besten zueinander passenden Bereiche von zwei Sequenzen, jedoch sagt dies nicht unbedingt auch etwas über die biologische Signifikanz der gefundenen Zuordnung aus. Daher wird die Signifikanz einer Zuordnung, wie in `BLAST`, anhand des sogenannten *p–value* abgeschätzt, der angibt, wie wahrscheinlich es ist, den *Score* für die gefundene Zuordnung zufällig zu beobachten. Dieser *p–value* wird vom `BLAST`–Programm berechnet [175], kann aber auch für Alignments ohne Gaps einfach berechnet werden (siehe dazu [10]). Altschul und Gish [9] zeigen empirisch, daß die Berechnung des *p–values* auf lokale Alignments mit Gaps verallgemeinert werden kann und nicht notwendigerweise durch einen Vergleich gegen die *Scores* optimaler Alignments nicht verwandter Sequenzen [10] bestimmt werden muß.

7.2.1.2 Globale und *free–shift* Sequenzalignments als Orakel

Werden Sequenzalignments über die Methode der dynamischen Programmierung berechnet, so gibt es eine andere Möglichkeit, die Zuverlässigkeit verschiedener Bereiche im Alignment zu bewerten [227]. Dabei wird für eine bestimmte Zuordnung jeweils der *Score* für ein optimales Alignment, das diese Zuordnung enthält, in Beziehung zum *Score* eines optimalen Alignments gesetzt, das diese Zuordnung nicht enthält. Eine Position im Alignment ist dann umso zuverlässiger, je größer die Differenz der beiden *Scores* ist. Dieser positionelle Zuverlässigkeitsindex kann durch zweimalige dynamische Programmierung effizient berechnet werden, wobei eine zusätzliche *Score*–Matrix für die invertierten Sequenzen berechnet wird. Die zur Berechnung des Alignments benötigte Laufzeit verdoppelt sich daher nur, so daß die Zeitkomplexität von $O(n^2)$ für affine Gapkosten erhalten bleibt.
Dieses positionelle Zuverlässigkeitsmaß ist am sinnvollsten auf globale und *free–shift* Alignments anwendbar. In der RDP–Methode wird dieser Zuverlässigkeitsindex wie folgt zur Identifizierung von Bereichen genutzt, die mit überdurchschnittlicher Zuverlässigkeit einander zugeordnet werden können:

1. Berechne für das aktuelle Teilproblem das globale beziehungsweise *free–shift* Alignment und dessen zugehörigen Zuverlässigkeitsindex (siehe oben).
2. Das so annotierte Alignment wird von der RDP–Methode auf zwei Arten verwendet:
   (a) i. Bestimme zusammenhängende Bereiche, deren Zuverlässigkeitsindices einen gegebenen Grenzwert überschreiten und die länger als eine vorgegebene Mindestlänge (zum Beispiel > 3 Positionen) sind.
   ii. Benutze jede der so gefundenen lokalen Zuordnungen wie Ergebnisse anderer Orakel in der RDP–Methode.
   (b) i. Lösche die Zuordnungen, deren Zuverlässigkeitsindex den geforderten Mindestwert nicht erreicht.
   ii. Markiere alle danach nicht zugeordneten Bereiche als noch in späteren Phasen des Algorithmus zu lösen.
   iii. Verwende das so modifizierte Alignment wie Ergebnisse anderer Orakel.

Bei der Berechnung von globalen beziehungsweise *free–shift* Alignments wird die gleiche Kostenfunktion verwendet wie für die lokalen Alignments mit dem Unterschied, daß die Kosten für die Eröffnung von Gaps (α) und deren Verlängerung (β) nicht hoch, sondern angemessen für die jeweils verwendete Ähnlichkeitsmatrix gewählt werden (siehe dazu auch Abschnitt 5.1).
Im folgenden werden weitere Orakel vorgestellt, die nach Ähnlichkeiten suchen, die auf Sequenzebene nicht mehr erkennbar sind.

7.2.2 Orakel und Einkörperpotentiale

Wie bereits in Abschnitt 4.2.2 beschrieben, können alle Einkörperpotentiale in der Form sogenannter *Profile* kodiert werden. Das zugeordnete Optimierungsproblem kann optimal mit der Methode der dynamischen Programmierung gelöst werden. Damit können alle Einkörperpotentiale in der bereits für die sequenzbasierten Orakel beschriebenen Form (siehe Abschnitt 7.2.1) von der RDP–Methode als Orakel für die Bestimmung lokaler Zuordnungen genutzt werden. Die RDP–Methode benutzt Einkörperpotentiale, um lokale Zuordnungen zwischen Sequenzabschnitten und Strukturbereichen zu finden, die nicht aufgrund von Sequenzähnlichkeiten, sondern anhand struktureller Eigenschaften und Präferenzen von Aminosäuretypen für bestimmte strukturelle Umgebungen identifiziert werden können. Folgende Einkörperpotentiale können gegenwärtig von der RDP–Methode benutzt werden:

- Hydrophobizitätspotentiale (siehe Abschnitt 5.2.3.1):
  - Ein durch Komponentenzerlegung des Paarinteraktionspotentials in paarabhängige und paarunabhängige Komponenten nach Bryant und Lawrence berechnetes Hydrophobizitätspotential [48].
  - Ein aus der Voronoizerlegung abgeleitetes Hydrophobizitätspotential, das den Aminosäuretypen eine Präferenz für den dem Lösungsmittel zugänglichen Oberflächenanteil zuordnet [366].

  Ersteres wird in der Regel mit auf Distanzkontaktrelationen (siehe Definition 5.3) basierenden Paarpotentialen kombiniert, letzteres wird gemeinsam mit Paarpotentialen verwendet, die auf Voronoikontaktrelationen (siehe Definition 5.5) basieren.
- Kontaktkapazitätspotentiale (*CCP*), wie sie von der `123D`–Methode [5] zur Berechnung von Sequenzstrukturalignments benutzt werden. Es gibt folgende Varianten von Kontaktkapazitätspotentialen:
  - Kontaktkapazitätspotential (*CCP*),
  - bedingtes *CCP* (*CCCP*),
  - abstandsabhängiges *CCP* (*DCCP*),
  - winkelabhängiges *CCP* (*ACCP*).

  Eine detaillierte Beschreibung findet sich in Abschnitt 5.2.3.2. Zur Zeit wird in der Regel das Kontaktkapazitätspotential (*CCP*) verwendet.
- Ist eine Sekundärstrukturvorhersage für die Sequenz (zum Beispiel mittels `PHD` [288, 289]) berechnet worden, so wird das diese Vorhersage kodierende Profil gegen das Profil der Sekundärstruktur der Struktur (Zuordnung nach `DSSP` [173]) aliniert. Als Teillösungen werden die Bereiche des Alignments extrahiert, die weder Insertionen noch Mismatches enthalten.

Da Profile in der RDP–Methode wie Sequenzen durch dynamische Programmierung mit der untersuchten Sequenz verglichen werden, werden zur Bestimmung lokaler Zuordnungen die gleichen Verfahren angewendet, wie sie bereits für den Vergleich der Sequenzen beschrieben wurden (siehe Abschnitt 7.2.1). Der einzige Unterschied besteht darin, daß der Kostenfunktionsanteil $\phi^S(f', A', B')$ durch $\phi^P(f', A', B')$ ersetzt wird, der wie folgt definiert ist:

$$\phi^P(f', A', B') = \sum_{f'(i)\neq\emptyset} prof(a_i, env_B(f'(i)))$$

Die Funktion env_B beschreibt dabei die Umgebung einer Position in der Struktur B. Die von der Funktion env_B implementierte Form der Beschreibung hängt davon ab, welches Einkörperpotential in dem Orakel verwendet wird. Bei Kontaktkapazitätspotentialen ist dies zum Beispiel die Anzahl der unterschiedlichen Kontakte, die die Position in der Struktur hat, bei Hydrophobizitätspotentialen die Lösungsmittelzugänglichkeit und bei Sekundärstrukturvergleichen die Sekundärstruktur, in der sich diese Position befindet.
Die Funktion *prof* bewertet anhand der von *env* gelieferten Beschreibung, wie gut die Aminosäure a_i in diese Umgebung gemäß des Potentials paßt (siehe dazu auch Abschnitte 5.2.3.1 und 5.2.3.2). Falls das Orakel auf den vorhergesagten Sekundärstrukturen basiert, wird anstelle des Aminosäuretyps a_i die für diese Position der Sequenz vorhergesagte Sekundärstruktur verwendet und der Funktionswert der Funktion *prof* ist 1, falls Vorhersage und Sekundärstruktur an der Strukturposition übereinstimmen.
Wenn der Hydrophobizitätsanteil und der Kontaktkapazitätsanteil des Kostenfunktionbestandteils $\phi^P(f', A', B')$ getrennt betrachtet werden sollen, werden diese im folgenden mit $\phi^H(f', A', B')$ beziehungsweise $\phi^C(f', A', B')$ bezeichnet.
Auch mit dieser Art von Orakel werden in jeder Rekursion der RDP–Methode eine vorgegebene Anzahl möglicher lokaler Zuordnungen bestimmt. Daher treten auch die gleichen Probleme wie bei den Sequenzorakeln auf, so entstehen zum Beispiel auch hier identische, sehr ähnliche, nicht zulässige, kombinierbare und nicht kombinierbare Teillösungen.

7.2.3 Orakel und Mehrkörperpotentiale

Die Einbeziehung von Mehrkörperpotentialen beim Sequenzstrukturalignment bewirkt, daß das zu lösende Optimierungsproblem in die Komplexitätsklasse der *NP–vollständigen* Probleme fällt (siehe Abschnitt 4.2.3). Ursächlich ist dafür, daß bei der Bewertung einer Position im Alignment andere Positionen im Alignment bei der Auswertung des Zwei– oder Mehrkörperpotentials berücksichtigt werden müssen. In den meisten Sequenzstrukturalignmentverfahren ist aber zum jeweiligen Zeitpunkt nicht klar oder nicht vorhersagbar, welcher Aminosäurerest auf eine zu der aktuell betrachteten Position in Kontaktrelation stehende Position in der Struktur abgebildet werden wird (siehe Diskussion in Abschnitt 6.1).

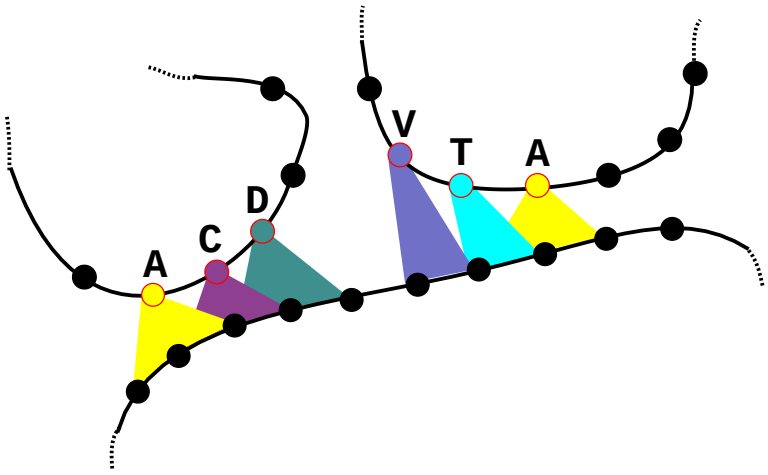


Abbildung 7.5: Zweikörperpotentiale und alinierte Positionen: Bereits alinierte Positionen (markiert mit Aminosäureresten) stehen in der Struktur mit noch nicht alinierten Positionen in Kontaktrelation. Die Farbkodierung der Kegel deutet die Aminosäuretypabhängigkeit der Mehrkörperpotentiale an.

In der RDP–Methode sind jedoch beim rekursiven Abstieg – außer direkt zu Beginn – bereits Teile der Abbildung der Sequenz in die Struktur bekannt. Diejenigen Positionen der Struktur, denen bereits ein Aminosäurerest der Sequenz zugeordnet wurde, stehen in Kontaktrelation mit Positionen der Struktur, die zu Teilproblemen gehören, die bislang nicht abgearbeitet wurden. Abbildung 7.5 zeigt, wie das Potential von bereits alinierten Positionen (oben, mit Aminosäurebezeichnern markiert) auf Positionen (unten) in der Struktur „abstrahlt“, die zu noch zu bearbeitenden Teilproblemen gehören.
Die Lösung eines aktuell bearbeiteten Teilproblems SP wird daher abhängig von bereits gefundenen Teillösungen, wenn Strukturpositionen von SP in Kontaktrelation zu in diesen Teillösungen zugeordneten Positionen stehen. Definition 7.4 formalisiert diese Abhängigkeit von bereits gefundenen Teillösungen.

Definition 7.4 (Kontaktabhängigkeit von Lösungen)
Sei $G_B^I = (V_B, E_B, \lambda, \mu)$ der wie in Definition 5.6 definierte Relationsgraph der Struktur B. Dann steht ein Teilproblem $SP \in \mathcal{T}$ in Kontakt *zu einem Teilproblem $SP' \in \mathcal{T}$, kurz $SP \sim_K SP'$, genau dann, wenn*

$$\exists\, i \in B|_{SP} :\ \exists\, j \in B|_{SP'} :\ (i,j) \in E_B$$

Sei f eine Lösung von Teilproblem SP, dann heißt die Lösung f' von SP' kontaktabhängig *von Teillösung f, kurz $f' \sim_K f$, genau dann, wenn $SP \sim_K SP'$.*

Wie bereits in Abschnitt 6.2 beschrieben, verfolgt die RDP–Methode auf jeder Rekursionsebene nicht nur eine erlaubte lokale Zuordnung sondern mehrere. Daher

enthält der Lösungsbaum $\mathcal{T}$ auch Teillösungen und Teilprobleme, die nicht mit einem bearbeiteten Teilproblem SP kombiniert werden können, da für ein Teilproblem weiter oben im Lösungsbaum durch verschiedene Orakel unterschiedliche Aufteilungen in Unterprobleme vorgegeben wurden.

Definition 7.5 (Verwandte (Teil–)Lösungen und Probleme)
P_x bezeichnet die Menge der Knoten, die auf dem direkten Weg von Knoten $x \in \mathcal{T}$ zur Wurzel des Lösungsbaums $\mathcal{T}$ liegen.
Zwei Knoten $x, y \in \mathcal{T}$ ((Teil–)Lösungen oder Probleme) heißen verwandt *genau dann, wenn*

- *$x \in P_y$ beziehungsweise $y \in P_x$ oder*
- *der erste gemeinsame Vorfahr von x und y in $\mathcal{T}$ ein Knoten aus $V_\wedge$ ist, der eine partielle Lösung repräsentiert.*

Die Menge $VL_x \subset \mathcal{H}$ ist die Menge der mit Knoten x verwandten Teillösungen und VP_x ist die Menge der mit Knoten x verwandten Teilprobleme.

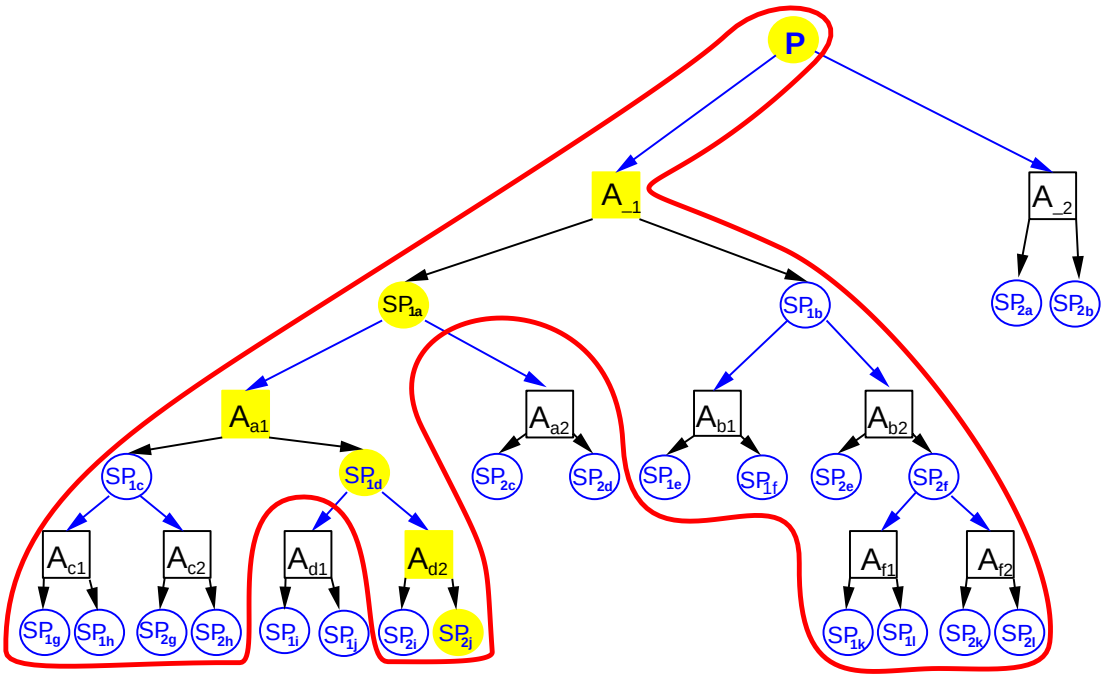


Abbildung 7.6: Abhängigkeitsbaum $\mathcal{T}$: Teillösungen und Teilprobleme sind wie in Abbildung 6.4 bezeichnet. Rot umrahmt ist der Teil des Lösungsbaums $\mathcal{T}$ von dem die Lösung des gelb markierten Teilproblems SP_{2j} abhängig sein kann. Gelb unterlegt sind die Knoten zum Weg der Wurzel des Lösungsbaums $\mathcal{T}$.

Abbildung 7.6 zeigt den Lösungsbaum $\mathcal{T}$, in dem von der RDP–Methode sowohl bereits gefundene Teillösungen (A_{ij}) als auch (noch offene) Teilprobleme (SP_{kl})

gespeichert werden. Die Wurzel (P) dieses Baumes entspricht dem vollständigen Sequenzstrukturalignmentproblem. Als Beispiel wird im folgenden das Teilproblem SP_{2j} des Teilbaumes $\mathcal{T}$ genauer betrachtet, welches in Abbildung 7.6 auf der untersten Hierarchiestufe des Lösungsbaums $\mathcal{T}$ gelb hinterlegt ist. Die mit Teilproblem SP_{2j} nach Definition 7.5 verwandten Teilprobleme und Teillösungen sind rot umrandet.
Bei den anderen Teilen des Lösungsbaums handelt es sich um Teilbäume, die Teilprobleme und Teillösungen enthalten, die nicht mit dem Teilproblem SP_{2j} kombiniert werden können. Diese Teile sind nach Definition 7.5 nicht verwandt, da die zugehörigen Teilbäume an Knoten von dem Pfad von SP_{2j} zur Wurzel abzweigen, die Teilproblemen entsprechen (in Abbildung 7.6 gelb unterlegt und mit SP beziehungsweise P gekennzeichnet).
Definition 7.6 schließt in Erweiterung von Definition 7.4 nicht verwandte Teillösungen aus der Liste der Abhängigkeiten aus, die bei der Lösung eines Teilproblems zu berücksichtigen sind.

Definition 7.6 (Abhängigkeit von Lösungen)
Die Lösung eines Teilproblems $SP \in \mathcal{T}$ heißt abhängig *von einer Teillösung f genau dann, wenn die Lösung kontaktabhängig von f ist und $f \in VL_{SP}$ gilt.*

Die Lösung des Teilproblems SP_{2j} in Abbildung 7.6 ist damit unter anderem unabhängig von den Teillösungen, die nicht aus $VL_{SP_{2j}}$ sind. In Abbildung 7.6 sind dies die Teillösungen, die in dem nicht rot umrandeten Teil des Lösungsbaums liegen.
Von allen anderen Teilen des Lösungsbaums kann die Lösung von SP_{2j} abhängig sein. In Abbildung 7.6 sind diese Teile des Lösungsbaums rot umrahmt. Ob die Lösung von SP_{2j} von Lösungen der verbleibenden Teilbäume abhängt oder nicht, kann nicht anhand der Eigenschaften des Lösungsbaums entschieden werden, sondern wird von den Kontaktrelationen vorgegeben, die durch die zugrundeliegende Proteinstruktur festgelegt sind. Wenn Strukturbereiche des Teilproblems SP_{2j} in Kontaktrelation zu Strukturbereichen stehen, die Bestandteil eines der verbleibenden Teilprobleme sind, dann ist die Lösung von SP_{2j} abhängig von den Lösungen des betreffenden Teilproblems, da diese die Kontaktpartner für Positionen in SP_{2j} festlegen.
Die RDP–Methode verfolgt auf jeder Rekursionsebene nicht nur *eine* erlaubte lokale Zuordnung sondern mehrere Alternativen. Für die Lösung eines Teilproblems bedeutet dies, daß durch die bereits gefundenen Teillösungen, von denen die Lösung des Teilproblems nach Definition 7.6 abhängig ist, nicht nur eindeutige Zuordnungen von Aminosäureresten der Sequenz zu Kontaktpartnern des aktuell bearbeiteten Teilproblems vorgegeben werden, sondern zum Zeitpunkt der Lösung des Teilproblems noch alternative Belegungen der Kontaktpartner möglich sind. Die Definition 7.7 trennt die bei der Bearbeitung eines Teilproblems zuberücksichtigenden Strukturpositionen aus B entsprechend auf:

Definition 7.7 (Teilmengen abhängiger Strukturpositionen)
Die Teilmenge B^1_{SP} enthält die Strukturpositionen aus B, denen in Bezug auf ein bearbeitetes Teilproblem SP bereits eindeutig ein *Aminosäurerest der Sequenz zugeordnet wurde:*

$$B^1_{SP} = \left\{ i \in B \backslash B|_{SP} \mid \exists f \in VL_{SP} : \; f^{-1}(i) \neq \emptyset \; \wedge \forall g \in VL_{SP} : \; g^{-1}(i) = f^{-1}(i) \right\}$$

Die Teilmenge B^*_{SP} enthält die Strukturpositionen aus B, denen bisher in verschiedenen Alternativen, die mit dem Teilproblem SP kombiniert werden können, unterschiedliche *Aminosäuren zugeordnet wurden:*

$$B^*_{SP} = \left\{ i \in B \backslash B|_{SP} \mid \exists f, g \in VL_{SP} : \; f^{-1}(i) \neq g^{-1}(i) \right\}$$

Die Teilmenge $B^\emptyset_{SP}$ enthält die Strukturpositionen aus B, denen bisher noch kein *Aminosäurerest der Sequenz zugeordnet wurde:*

$$B^\emptyset_{SP} = \left\{ i \in B \backslash B|_{SP} \mid \forall f \in VL_{SP} : \; f^{-1}(i) = \emptyset \right\}$$

Zu welcher der drei Mengen B^1_{SP}, B^*_{SP} oder $B^\emptyset_{SP}$ eine Strukturposition gehört, hängt wesentlich von der Lage des Alignments im Lösungsbaum $\mathcal{T}$ ab, das die Zuordnung für diese Position bestimmt.

Lemma 7.1
Sei $SP \in \mathcal{T}$ und $i \in B \backslash B|_{SP}$, dann gilt:

$$\exists f \in P_{SP} : \; f^{-1}(i) \neq \emptyset \implies i \in B^1_{SP}$$

Die Umkehrfolgerung gilt im Allgemeinen nicht.

Lemma 7.1 besagt, daß Strukturpositionen, denen Aminosäurereste der Sequenz in partiellen Lösungen zugeordnet werden, die auf dem direkten Weg vom bearbeiteten Teilproblem SP zur Wurzel des Lösungbaums $\mathcal{T}$ liegen (also nach Definition 7.5 Elemente von P_{SP} sind), zur Menge B^1_{SP} gehören. In Abbildung 7.6 gehören so zum Beispiel alle Strukturpositionen zu $B^1_{SP_{2j}}$, die von gelb hinterlegten partiellen Lösungen mit Sequenzpositionen aliniert werden.
Der Beweis des Lemmas folgt aus der Konstruktion der Teilprobleme (siehe Abschnitt 7.4), die auf dem Weg von der Wurzel zum bearbeiteten Teilproblem SP liegen. Wird eine Strukturposition innerhalb einer partiellen Lösung zugeordnet, so ist sie nicht mehr Bestandteil der aus dieser Zuordnung resultierenden Teilprobleme und kann somit nicht mehr durch andere partielle Lösungen anderweitig aliniert werden, die mit dem Teilproblem SP gemäß Definition 7.5 verwandt sind.
Die Umkehrfolgerung aus Lemma 7.1 gilt nicht, da $i \in B^1_{SP}$ auch dann gilt, wenn $\forall f, g \in VL_{SP} \backslash P_{SP} : \; f^{-1}(i) = g^{-1}(i)$.
Für Kontaktbereiche eines Teilproblems SP, denen durch Teillösungen $f \notin P_{SP}$ Aminosäurereste der Sequenz zugeordnet werden, kann die Zuordnung eindeutig

oder mehrdeutig sein. Ob die Zuordnung aus Sicht von SP eindeutig ist oder nicht, hängt von der Probleminstanz und dem aktuellen Zustand des Lösungsbaums $\mathcal{T}$ ab. Für diese Bereiche kann mit der Definition der Mengen B^1_{SP} und B^*_{SP} nur die folgende Aussage getroffen werden:

Lemma 7.2
Sei $SP \in \mathcal{T}$ und $i \in B \backslash B|_{SP}$, dann gilt folgende Äquivalenz:

$$\exists\ f \in VL_{SP} :\ f^{-1}(i) \neq \emptyset \iff i \in B^1_{SP} \cup B^*_{SP}$$

Angenommen für das Teilproblem SP_{2j} aus Abbildung 7.6 gilt $SP_{2j} \sim_K SP_{1b}$. Dann ist es zum Beispiel möglich, daß einer Position im durch SP_{1b} definierten Teilbaum durch das Alignment A_{b1} eine andere Aminosäure zugeordnet wird als durch das Alignment A_{b2}. Da es sich bei diesen Teillösungen um gleichberechtigte Alternativen zur Lösung des Teilproblems SP_{1b} handelt, kann diese Mehrdeutigkeit des möglichen Kontaktpartners erst aufgelöst werden, wenn die Entscheidung für eine dieser Alternativen im Teilproblem SP_{1b} getroffen wurde.
Für Strukturpositionen eines betrachteten Teilproblems SP, die kontaktabhängig von einem noch nicht bearbeiteten Teilproblem sind, kann in der Regel keine Aussage über den oder die Partner einer Kontaktrelation getroffen werden.
Natürlich ist die Zuordnung von Strukturpositionen zu einer der drei Mengen B^1_{SP}, B^*_{SP} und $B^\emptyset_{SP}$ dynamisch und hängt stark vom Fortschreiten der Berechnung und damit dem Zeitpunkt ab, zu dem die Lösung des Teilproblems SP von der RDP–Methode angegangen wird. Je größer die Menge B^1_{SP} oder eventuell auch die Menge B^*_{SP} ist, desto mehr Kontakte können anhand des resultierenden Strukturmodells bewertet werden.
Daher stellt in der RDP–Methode neben der Größe der Teilprobleme die Mächtigkeit insbesondere der Menge B^1_{SP} mögliches Kriterium für die Sortierung der Prioritätswarteschlange der offenen Teilprobleme und damit für die Festlegung der Abarbeitungreihenfolge der Teilprobleme dar.
Die RDP–Methode versucht bei der Lösung von Teilproblemen so viel Informationen über die Kontaktpartner auszunutzen, wie zum Zeitpunkt verfügbar ist, an dem die Lösung eines Teilproblems ansteht. Dazu werden bei der Lösung eines Teilproblems SP verschiedene Strategien verfolgt, die in den folgenden Abschnitten detaillierter beschrieben werden:

Orakel mit eindeutig bestimmten Kontaktpartnern: Die Suche wird auf die Bereiche der Struktur eingeschränkt, für die hinreichend viele Kontaktpartner in B^1_{SP} und damit bereits eindeutig festgelegt sind. Für diese Teile der Struktur können Kontaktrelationen anhand der neuen Kontaktpartner bewertet werden (Abschnitt 7.2.3.1).

Orakel mit alternativen Kontaktpartnern: Bei der Suche werden die nicht eindeutigen Zuordnungen der Kontaktpartner aus B^*_{SP} explizit als alternative Zuordnungen berücksichtigt. Dies führt zu einer Aufspaltung in verschiedene Orakel (Abschnitt 7.2.3.2).

Orakel mit gemittelten Kontaktpartnern: Die alternativen Zuordnungen für Kontaktpartner aus B^*_{SP} werden nicht explizit in verschiedenen Orakelschritten modelliert, sondern gemittelt betrachtet (Abschnitt 7.2.3.3).

Orakel mit *semi–eingefrorener* Approximation: Für die Kontaktpartner aus B^1_{SP} werden die zugeordneten Aminosäuretypen aus der Sequenz bei der Berechnung lokaler Zuordnungen zur Bewertung der Kontaktrelationen verwendet, und für Kontakte zu Positionen aus $B^\emptyset_{SP}$ werden die Aminosäuretypen aus der Struktur übernommen (Abschnitt 7.2.3.4).

Orakel mit anderen Kostenfunktionsanteilen: Es werden Kontaktrelationen mit Partnern aus B^1_{SP} (und B^*_{SP}) bewertet, aber zusätzlich werden Kostenfunktionsbestandteile aus den anderen vorgestellten Orakeltypen verwendet (Abschnitt 7.2.3.5).

Dies sind die wesentlichen Arten, wie Kontaktrelationen von der RDP–Methode bei der Berechnung lokaler Zuordnungen eingesetzt werden. Da aber eine wesentliche Idee der RDP–Methode darin besteht, alle Kostenfunktionsbestandteile so zu modellieren, daß sie auch kombiniert verwendbar sind, sind auch Mischformen möglich.

7.2.3.1 Orakel mit eindeutig bestimmten Kontaktpartnern

Für ein betrachtetes Teilproblem SP sind mit Lemma 7.1 zunächst einmal die Kontaktpartner eindeutig bestimmt (aus B^1_{SP}), denen durch Teillösungen $f \in P_{SP}$ Reste der Sequenz A zugeordnet werden. Es gibt zusätzlich aber auch Kontaktpartner die dadurch zu B^1_{SP} gehören, daß ihre Belegung durch Aminosäurereste aus der Sequenz für alle alternativen Teillösungen identisch sind.
Das Orakel mit eindeutig bestimmten Kontaktpartnern sucht nach Teillösungen für das Teilproblem SP, die für die Aminosäurebelegung der Kontaktpartner aus B^1_{SP} optimal sind. Die Teillösungen werden in folgenden Schritten berechnet.

1. Bestimme die Kontaktrelationen zu Strukturpositionen aus B^1_{SP}.
2. Markiere Strukturpositionen des aktuellen Teilproblems als *verwendbar*, falls zum Beispiel für mehr als ein Viertel ihrer Kontaktrelationen der oder die Kontaktpartner aus B^1_{SP} sind.
3. Als Sonden werden die Bereiche in der Struktur bezeichnet, für die alle Strukturpositionen als verwendbar markiert sind und die eine gewisse Mindestlänge (zum Beispiel > 3 Positionen) haben.
4. Suche für jede dieser Sonden in der Sequenz des Teilproblems nach optimalen Belegungen gemäß des Mehrkörperpotentials.

Die Kontaktrelationen, an denen Strukturpositionen des aktuellen Teilproblems beteiligt sind, können über den Indexvektor des Teilproblems aus der Strukturbe-

schreibung der Gesamtstruktur übernommen werden. Die Bestimmung der eindeutig festgelegten Kontaktpartner erfolgt in zwei Schritten. Zunächst werden in einer Liste die Positionen aus B^1_{SP} gesammelt, indem die zu Teilproblem SP verwandten Teillösungen VL_{SP} im Lösungsbaum durchlaufen werden. Im nächsten Schritt wird die Liste in einen Index umgesetzt, über den für jede Strukturposition auf die eindeutig alinierte Sequenzposition zugegriffen werden kann. Durch diese Indirektion kann nicht nur auf Sequenzen sondern auch auf Profile von Sequenzen zugegriffen werden, was sowohl für die Erweiterung auf multiples Sequenzstrukturalignment als auch für die Berücksichtigung nicht eindeutig alinierter Strukturpositionen genutzt wird (siehe Abschnitt 7.2.3.3).
Wird ein Teilproblem bearbeitet, während die Abarbeitung durch die RDP–Methode in Tiefensuche erfolgt, kann der Aufbau dieses Indexes inkrementell erfolgen, da sich bei der Abarbeitung der Teilprobleme in Tiefensuche nur lokale Änderungen im Lösungsbaum ergeben. Dabei wird zum einen ausgenutzt, daß für Teilprobleme, die durch das gleiche Teilproblem entstanden sind, auch die gleichen Kontaktpartner gültig sind, so daß die Berechnung der Kontaktpartner nur einmal durchgeführt werden muß. Außerdem haben in diesem Fall alle Unterprobleme eines Teilproblems die gleichen Kontaktpartner, bis auf die Kontaktpartner, die durch die das jeweilige Teilproblem definierende Teillösung festgelegt werden. Daher muß in Tiefensuchephasen der RDP–Methode nur einmal der in Abbildung 7.6 rot umrahmte Bereich durchsucht werden. Nachdem dies einmal erfolgt ist, muß höchstens noch der Bereich aktualisiert werden, der die in verwandten Teillösungen auf der gleichen oder nächsthöheren Hierarchiestufe festgelegten Kontaktpartner speichert.
Werden die Teilprobleme von der RDP–Methode in Breitensuche durchlaufen, sind derartige Vereinfachungen nicht möglich, da Änderungen im Lösungsbaum nicht ausgeschlossen werden können, die die Kontaktpartnerbelegung für ein aktuell bearbeitetes Problem ändern.
Die Schritte 2. Markierung und 3. Identifizierung und Auswahl sind mit dieser Vorbearbeitung sehr einfach realisierbar. Jede so bestimmte Sonde definiert dann ein Unterproblem, das durch den durch die Sonde bestimmten Ausschnitt aus der Struktur und den Sequenzabschnitt des bearbeiteten Teilproblems beschrieben ist.
Diese Sonden sind so bestimmt, daß für viele der Kontaktrelationen die neuen Partner aus der Sequenz bekannt sind. Daher ist es sinnvoll, für diese Sonden nach Teilen aus der Sequenz zu suchen, die bezüglich der Kontaktrelationen gut auf diese Positionen in der Struktur abgebildet werden können. Die bekannten Kontaktrelationspartner induzieren, wie in Abbildung 7.5 gezeigt, für die Positionen der Sonde Präferenzen für bestimmte Aminosäuren. Diese Präferenzen geben also für jede Position und jeden Aminosäuretyp an, wie gut dieser Aminosäuretyp an die jeweilige Position gemäß der Kontaktrelationen, der Kontaktrelationspartner und gemäß des verwendeten Mehrkörperpotentials E_{pot} paßt. Zur Vereinfachung der Schreibweise wird im folgenden davon ausgegangen, daß das Mehrkörperpo-

tential E_{pot} auf Paarkontaktrelationen definiert ist.
Sei $A' = A|_{SP}$, B' eine derartige Sonde und $G_B^I = (V_B, E_B, \lambda, \mu)$ der Kontaktrelationsgraph der Struktur B. Weiterhin bezeichne g_j jeweils die Teillösung, die der Strukturposition $j \in B_{SP}^1$ eine Aminosäure aus $A \backslash A|_{SP}$ eindeutig zuordnet. Bezeichne $\mathcal{G}^1(j)$ die Menge mit der Teillösung g, die einer Position $j \in B_{SP}^1$ eine Aminosäure aus Sequenz A eindeutig zuordnet. Der Kontaktpotentialanteil $\phi_{pot}^1(f', A', B')$ der für die Bestimmung der Teillösung f' zu optimierenden Kostenfunktion $\phi(f', A', B')$ kann dann wie folgt geschrieben werden:

$$\phi_{pot}^1(f', A', B') = \sum_{f'^{-1}(i) \neq \emptyset} \sum_{((i,j),g) \in (E_B \times \mathcal{G}^1(j))} E_{pot}\Big(a_{f'^{-1}(i)},\ a_{g^{-1}(j)},\ \lambda(i,j)\Big)$$

Das Potential E_{pot} bewertet dabei die Kontaktrelation $(i,j) \in E_B$ anhand der Aminosäuretypen $a_{f'^{-1}(i)}$ und $a_{g^{-1}(i)}$ und der Attribute des Kontaktes $\lambda(i,j)$. Seien die affinen Gapkosten, wie in Abschnitt 7.2.1 eingeführt, durch den Term $GAP(f', A', B')$ zusammengefaßt, dann ergibt sich die zu optimierende Kostenfunktion $\phi(f', A', B')$ zur Lösung des durch A' und B' definierten Teilproblems zu:

$$\phi(f', A', B') = \phi_{pot}^1(f', A', B') + GAP(f', A', B')$$

Die innere Summe des Kontaktpotentialanteils $\phi_{pot}^1(f', A', B')$ kann bei der Berechnung der optimalen Teillösung f', wie in Abbildung 7.7 angedeutet, in der Form eines Kontaktpotentialprofils (KPP) aufbereitet werden. Ein Eintrag k, l in dem Profil $KPP_{B'}$ des Teilproblems SP sagt also, wie gut ein Aminosäuretyp $k \in A'$ an einer Position l zu den von der Struktur vorgegebenen und durch die bisher gefundenen Teillösungen teilweise mit Aminosäuren aus der Sequenz belegten Kontaktrelationen paßt.
Dieses Profil kodiert somit alle Kontaktrelationen zu bereits eindeutig alinierten Positionen, somit entspricht das optimale Alignment von A' mit dem Profil $KPP_{B'}$ der optimalen Zuordnung von Teilen der Sequenz zu den Sondenpositionen unter den durch die bisherigen Teillösungen vorgegebenen Randbedingungen und gemäß der Kontaktrelationen und dem Mehrkörperpotential. Das optimale Alignment einer Sequenz gegen ein Kontaktpotentialprofil ist mit Methoden der dynamischen Programmierung optimal und effizient berechenbar.
In dieser Form des Orakels tritt jedoch häufig das Problem auf, daß über die Sonden nur unsignifikante Teillösungen gefunden werden, da nicht genügend Kontaktpartner eindeutig festgelegt sind. Zur Umgehung dieses Problems werden in den im folgenden vorgestellten Orakeln folgende Erweiterungen vorgenommen:

- Es werden auch nicht eindeutig alinierte Positionen benutzt.
- Falls ein Kontaktpartner noch nicht aliniert wurde, wird der Aminosäuretyp aus der Struktur übernommen.
- Die Sonden werden auch auf nicht zusammenhängende Bereiche ausgedehnt und in Sonden sind auch Positionen ohne bekannte Kontaktpartner erlaubt.

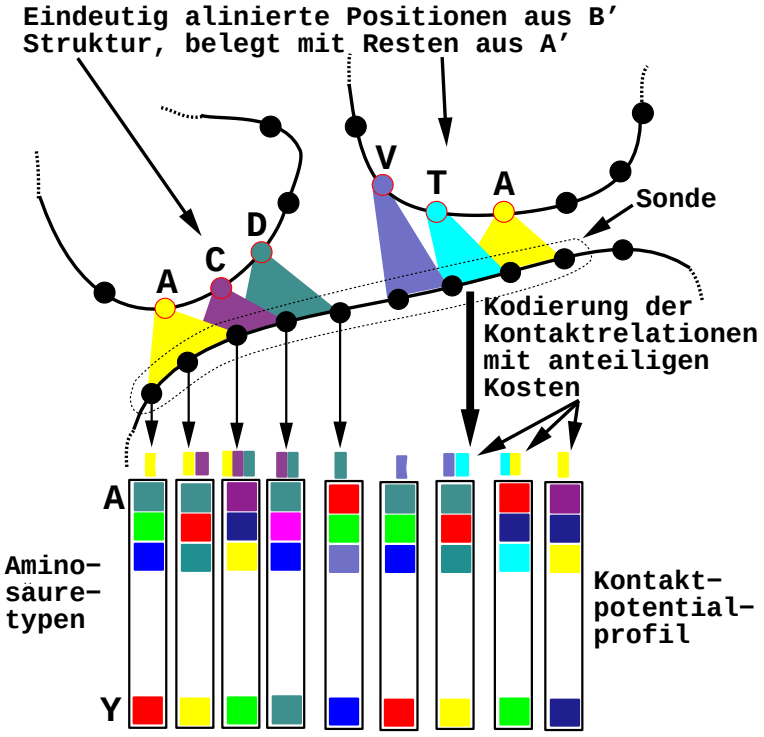


Abbildung 7.7: Kodierung von Kontaktrelationen als Profil: Positionen aus B^1_{SP} stehen in Kontaktrelation mit Positionen der Sonde B'. Die Wirkung des Mehrkörperpotential über die Kontaktrelationen auf die Positionen aus B' wird in ein Kontaktpotentialprofil übersetzt.

- In der Kostenfunktion werden auch Bestandteile der in den vorangegangenen Kapiteln vorgestellten Orakeltypen verwendet.

7.2.3.2 Orakel mit alternativen Kontaktpartnern

Sollen zur Lösung des Teilproblems SP nicht Kontakte mit Kontaktpartner aus B^1_{SP} sondern auch aus B^*_{SP} berücksichtigt werden, so werden die verschiedenen alternativen Zuordnungen sequentiell abgearbeitet. Abbildung 7.8 zeigt, wie dies mit einer leichten Modifizierung der für das Orakel mit eindeutigen Kontaktpartnern verwendet Kostenfunktion realisiert werden kann. Die Modifikation besteht darin, daß die innere Summe nicht über die $j \in B^1_{SP}$ läuft, sondern über die $j \in B^{1*}_{SP}$, wobei B^{1*}_{SP} jeweils die für die $j \in B^*_{SP}$ um eine der möglichen Zuordnungen erweiterte Menge B^1_{SP} ist. Dies ist jedoch nur dann möglich und sinnvoll, wenn die Anzahl der alternativen Belegungen klein ist oder wenn für jede der alternativen Belegungen der Kontaktpartner die Teillösung gleich oder zumindest ähnlich ist, so daß die Anzahl der unterschiedlichen Alternativen für ein Teilproblem beschränkt bleibt oder durch Zusammenfassung ähnlicher Teillösungen

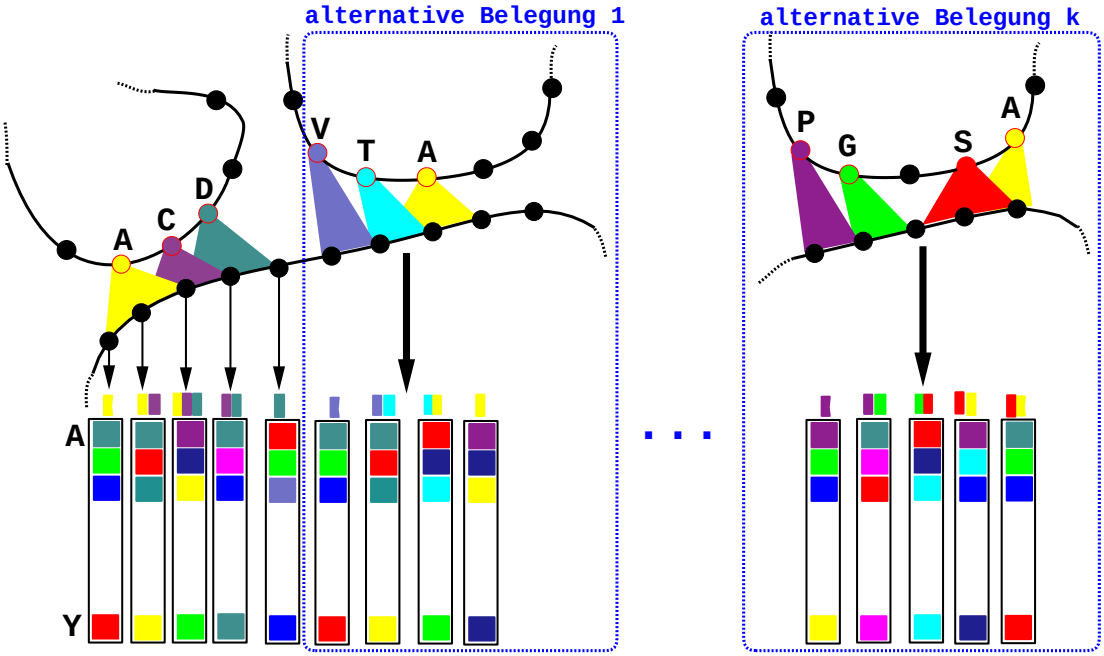


Abbildung 7.8: Mehrkörperpotential–Orakel mit alternativen Belegungen der Kontaktpartner: Es werden nur die von der alternativen Belegung betroffenen Bereiche des Kontaktpotentialprofils aktualisiert.

eingeschränkt werden kann.
Sofern sich die alternativen Belegungen der Kontaktpartner nur für wenige Positionen oder eingrenzbare Bereiche der Struktur unterscheiden, kann bei der Berechnung der Teillösungen die inkrementelle Aufbaustrategie ausgenutzt werden, wie sie im vorangegangenen Abschnitt vorgestellt wurde. Dabei wird zwischen den verschiedenen Aufrufen des Orakels der variable Teil aus dem Kontaktpotentialprofil herausgerechnet und durch den Anteil ersetzt, der aus den alternativen Belegungen der nächsten Alternative entsteht.
Das Problem bei der Verwendung eines derartigen Orakels besteht nicht in der Berechnung der Lösungen, sondern in deren Vielfalt und darin, daß ihre Optimalität an Teilen der Lösung hängt, deren Zuordnung bis zu diesem Zeitpunkt gerade nicht eindeutig möglich war und daher als vage anzusehen ist.

7.2.3.3 Orakel mit gemittelten Kontaktpartnern

Eine detaillierte Bewertung von Kontakten, an denen Kontaktpartner beteiligt sind, denen bisher keine Aminosäure aus der Sequenz eindeutig zugeordnet wurde, kann zum Zeitpunkt des Orakelaufrufs nicht erfolgen, beziehungsweise ist abhängig von der jeweils zugrunde liegenden Belegungsalternative. Daher ist es sinnvoll, nicht die verschiedenen alternativen Belegungen getrennt zu betrachten,

sondern die verschiedenen Alternativen zusammenzufassen.
Dazu wird statt der Präferenzwerte zu einem eindeutigen Kontaktpartner der Durchschnittswert der Potentialwerte aller Alternativkontaktpartner in dem Kontaktpotentialprofil *KPP* für ein Teilproblem *SP* verwendet. Bezeichne $\mathcal{G}^*(j)$ die Menge der alternativen Teillösungen, die einer Position $j \in B^*_{SP}$ eine Aminosäure aus der Sequenz A zuordnen. Mit den aus Abschnitt 7.2.3.1 bekannten Namenskonventionen ergibt sich der Kostenfunktionsanteil $\phi^*_{pot}(f', A', B')$, der Kontaktrelationen von Strukturpositionen aus SP mit Positionen $j \in B^*_S$ bewertet, zu:

$$\phi^*_{pot}(f', A', B') = \sum_{f'^{-1}(i) \neq \emptyset} \sum_{(i,j) \in E_B} \frac{\sum_{g \in \mathcal{G}^*(j)} E\left(a_{f'^{-1}(i)},\ a_{g^{-1}(j)},\ \lambda(i,j)\right)}{|\mathcal{G}^*(j)|}$$

Kontaktrelationen zu Positionen $j \in B^1_S$ werden, wie in Abschnitt 7.2.3.1, mit der Funktion $\phi^1_{pot}(f', A', B')$ bewertet, so daß in diesem Orakel folgende Kostenfunktion verwendet wird:

$$\phi(f', A', B') = \phi^1_{pot}(f', A', B') + \phi^*_{pot}(f', A', B') + GAP(f', A', B')$$

Der Vorteil dieser Vorgehensweise ist, daß die Sondengrößen wesentlich zunehmen und trotzdem die Anzahl der von diesem Orakel generierten Lösung überschaubar bleibt. Der Nachteil besteht darin, daß nicht in allen Fällen die endgültigen Kontaktpartner bei der Bewertung einer Kontaktrelation, sondern nur der Durchschnitt über die möglichen Kontaktpartner verwendet wird.

7.2.3.4 Orakel mit *semi–eingefrorener* Approximation

Bei der von vielen Autoren [100, 118, 321, 365] zum Sequenzstrukturalignment mit Mehrkörperpotentialen verwendeten eingefrorenen Approximation (siehe auch Abschnitt 4.2.3.2) werden anstelle der wirklichen Kontaktpartner in der zu berechnenden Modellstruktur als Näherung die Aminosäuretypen aus der bekannten Struktur übernommen. Die Anpassung auf die wirklichen Kontaktpartner in dem Modell erfolgt in iterativen Alignmentschritten.
Aufgrund der RDP–Strategie, zunächst die signifikanten Zuordnungen zu berechnen und nicht, wie bei der dynamischen Programmierung, am Anfang von Sequenz und Struktur zu beginnen, sind häufig, wenn eine Lösung für ein Teilproblem gesucht wird, bereits die wesentlichen Kontaktpartner im späteren Modell bekannt. Daher werden Kontakte mit Kontaktpartner aus B^1_{SP} durch den aus Abschnitt 7.2.3.1 bekannten Kostenfunktionsterm $\phi^1_{pot}(f', A', B')$ bewertet.
Für Kontaktrelationen mit Partnern aus $B^\emptyset_{SP}$ und B^*_{SP} wird im Sinne der eingefrorenen Approximation davon ausgegangen, daß durch die spätere Zuordnung von Aminosäuren die wesentlichen Kontakteigenschaften erhalten bleiben.

$$\phi^{fa}_{pot}(f', A', B') = \sum_{f'^{-1}(i) \neq \emptyset} \sum_{(i,j) \in E_B \ \wedge \ j \in B^\emptyset_{SP} \cup B^*_{SP}} E\left(a_{f'^{-1}(i)},\ b_j,\ \lambda(i,j)\right)$$

Als Sonde im Sinne der in den vorangegangenen Abschnitten beschriebenen Orakel kann damit die gesamte Teilstruktur des aktuellen Teilproblems verwendet werden und die zu optimierende Kostenfunktion ergibt sich zu:

$$\phi(f', A', B') = \phi^1_{pot}(f', A', B') + \phi^{fa}_{pot}(f', A', B') + GAP(f', A', B')$$

Obwohl die Sonde die gesamte Teilstruktur umfaßt, wird auch in Orakeln mit der sogenannten *semi–eingefrorenen* Approximation zur Bewertung von Kontaktrelationen mit den Methoden des lokalen Alignments nach lokal gut passenden Bereichen gesucht.
Es ist klar, daß die wirklich im späteren Modell interagierenden Aminosäuren desto stärker in die Kontaktpotentialprofile einfließen, je mehr von der RDP–Methode bereits eindeutig zugeordnet werden konnte. Da dieses Verhalten ursächlich auf die Lösungsstrategie der RDP–Methode zurückgeführt werden kann, adaptiert sich die RDP–Methode an die jeweils verfügbaren Informationen über das spätere Strukturmodell.
Ein weiterer Vorteil der semi–eingefrorenen Approximation ist, daß sie auch dann anwendbar ist, wenn die bereits eindeutig zugeordneten Strukturpositionen für ein aktuell bearbeitetes Teilproblem nicht ausreichen, um entsprechend spezifische Sonden zu definieren.

7.2.3.5 Orakel mit anderen Kostenfunktionsanteilen

Eine andere Möglichkeit, mangelndes Wissen über die zukünftigen Kontaktpartner auszugleichen, besteht darin, ein Orakel nicht allein auf dem Kontaktpotential aufzubauen. In den Orakeln mit anderen Kostenfunktionsanteilen wird daher nach Teilen der Sequenz gesucht, die nicht nur eine optimale Belegung der Kontaktrelationen darstellen, sondern gleichzeitig auch bezüglich der in den zuvor vorgestellten Orakeln optimal sind.
Dies ist in der RDP–Methode einfach möglich, da alle Kostenfunktionen in uniformer Weise entweder durch eine Austauschmatrix oder durch Profile beschrieben werden. Werden also zum Beispiel Kontaktkapazitätspotentiale und Kontaktpotentiale gleichzeitig als Optimierungskriterien benutzt, so sind nur die jeweiligen Kostenfunktionsanteile mit einer geeigneten Gewichtung zu kombinieren.

$$\begin{aligned}\phi(f', A', B') \;=\; & \gamma * \phi^S(f', A', B') \;+\; \delta * \phi^C(f', A', B') \;+ \\ & \epsilon * \phi^H(f', A', B') \;+\; \zeta * \phi^1_{pot}(f', A', B') \;+ \\ & GAP(f', A', B')\end{aligned}$$

Auch in dem Orakel mit kombinierter Kostenfunktion wird, wie im vorangegangenen Orakel, mit den Methoden des lokalen Alignments nach gut zueinander passenden Bereichen in der gesamten Sequenz und der gesamten Struktur des Teilproblems gesucht. Die Gewichtungen γ, δ, ϵ und ζ sind empirisch ermittelt (siehe Abschnitt 7.7) und entsprechen denen, die auch bei der Auswahl von Teillösungen verwendet werden (siehe Abschnitt 7.6.2).

7.2.3.6 Kombinations– und Spezialorakel

Die in den vorangegangenen Abschnitten vorgestellten Orakel stellen nur eine Auswahl der in der RDP–Methode eingesetzten Orakel dar. Zum Beispiel sind nahezu alle vorgestellten Orakeltypen miteinander kombinierbar. Aber auch neuartige Orakel können einfach in die RDP–Methode integriert werden. Dies ist dann sinnvoll, wenn spezielle Sequenzstrukturalignments für spezielle Anwendungsbeispiele berechnet werden sollen, für die zum Beispiel besondere Randbedingungen vom Anwender vorgegeben werden, die in Orakel umgesetzt werden können, die mit den zuvor beschriebenen Mechanismen nur unzureichend modellierbar sind.
Obwohl alle an einem Knoten weiterverfolgten Teillösungen unterschiedlich sind, entstehen im Lösungsbaum dennoch Unterprobleme, die die gleichen Sequenz– und Strukturbereiche beinhalten. Während die in Abschnitt 7.2.3 vorgestellten Orakel von anderen zum jeweiligen Zeitpunkt bereits berechneten Teillösungen dynamisch abhängen und so jeweils neu berechnet werden müssen, ist dies für die in den Abschnitten 7.2.1 und 7.2.2 vorgestellten Sequenz– und Profilorakel nicht der Fall. Daher können die in diesem Orakeln für ein Teilproblem bestimmten Teillösungen direkt für später bearbeitete Teilprobleme wiederverwertet werden, die die gleichen Sequenz– und Strukturbereiche beinhalten, und müssen nicht mehrfach mit quadratischem Aufwand neu berechnet werden.
Von der RDP–Methode werden bereits berechnete, von externen Kontakten unabhängige Teillösungen so zwischengespeichert, daß auf sie indiziert über den Anfangspunkt des Teilproblems in der Struktur und beschrieben durch die weiteren Anfangs– und Endpunkte effizient zugegriffen werden kann. In zahlreichen Fällen kann so der Berechnungsaufwand für eine Teillösung von quadratisch auf linear reduziert werden, da nur noch das Erstellen einer Kopie der Teillösung für die weiteren Schritte der RDP–Methode notwendig ist. In typischen Anwendungsbeispielen kann so mehr als ein Viertel der Laufzeit eingespart werden.

7.2.4 Orakel und Informationen über aktive Stellen

Mögliche Bindungsstellen können sowohl aus Mutationsexperimenten und Aktivitätstest als auch aus Datenbanken (zum Beispiel `PROSITE` [17, 20]) abgeleitet werden. Die folgenden Abschnitte zeigen, daß diese Art von Information nicht nur als Kriterium für die Auswahl zulässiger Teillösungen und das Aussortieren nicht zulässiger Teillösungen (siehe Abschnitt 7.6.2.3) verwendet werden kann, sondern auch zur direkten Generierung möglicher Lösungen für ein Teilproblem.

7.2.4.1 Orakel und Information aus Experimenten

Stammt die Bindungsstelleninformation aus dem Wissen über die Funktion des Proteins und aus Mutationsexperimenten, so sind die an der Ausbildung der aktiven Stelle beteiligten Aminosäurereste in der untersuchten Sequenz in der

Regel bekannt. Dann sind vier Fälle für die Sequenz A und die Struktur B zu unterscheiden:

1. Die Proteine A und B haben vergleichbare Funktionen und die Zuordnung der funktionell relevanten Aminosäurereste von A auf entsprechende Positionen von B ist
   (a) bekannt und eindeutig, oder
   (b) nicht eindeutig beziehungsweise nicht möglich.
2. Für die Proteine A und B sind bislang nur unvergleichbare oder keine Funktionen bekannt und
   (a) ein funktionelles Sequenzmuster oder ein entsprechender regulärer Ausdruck paßt zu einem oder mehreren Sequenzsegmenten von B, oder
   (b) auf Sequenzebene kann kein Hinweis auf eine vergleichbare Funktion von B gefunden werden.

Der Fall 1(a) tritt typischerweise dann auf, wenn das Ziel die Berechnung eines genauen Sequenzstrukturalignments zur Verwendung in der vergleichenden Modellierung (siehe Abschnitt 4.1) ist und die Funktion beider Proteine im wesentlichen bekannt ist. In diesem Fall werden die Reste der aktiven Stelle von A auf die zugehörigen Positionen in der Struktur von B abgebildet. Im Sinne der RDP–Methode handelt es sich bei dieser Zuordnung um eine mögliche Ausprägung des *Orakels*.
Da es sich dabei um eine klare Zuordnung handelt, liefert diese Art von Orakel zusätzlich ein Kriterium für die Zulässigkeit der von anderen Orakeltypen generierten Teillösungen. Und zwar sind diese nur dann zulässig, wenn sie die vorgegebenen Zuordnungen enthalten oder noch offen lassen (siehe auch Abschnitt 7.6.2.3).
Wird die Zuordnung aktiver Stellen als Orakel benutzt, können als Ergebnis zwei Fälle eintreten:

1. Die aktive Stelle wird durch einen in der Sequenz zusammenhängenden Bereich von Aminosäureresten beschrieben (zum Beispiel die *P–Loop* in `ATP`–bindenden Proteinen (siehe Abschnitt 8.6)).
2. Die aktive Stelle wird durch Aminosäurereste gebildet, die in der Sequenz durch variable Teile von einander getrennt sind (zum Beispiel die Reste der katalytischen Triade in Serinproteasen).

Unabhängig davon, welcher der beiden Fälle vorliegt, kann das bearbeitete Problem anhand der zugeordneten Reste aus A und B in kleinere, noch zu lösende Unterprobleme aufgeteilt werden, deren Größe von den Positionen der beteiligten Reste in A und B abhängt. Der Unterschied besteht in der Anzahl der noch zu

lösenden Unterprobleme. Im Fall eines zusammenhängenden Bereiches entstehen zwei, im Allgemeinen bei n zusammenhängenden Bereichen $n+1$ Unterprobleme. Die Aufteilung in Unterprobleme wird in Abschnitt 7.4 beschrieben.
Im Fall, daß eine sinnvolle Zuordnung der Aminosäurereste der aktiven Stelle von A auf Positionen von B nicht (eindeutig) möglich ist (Fall 1(b)), kann die Bindungsstelleninformation nicht als Orakel, sondern nur als Randbedingung bei der Selektion zulässiger Teillösungen genutzt werden (siehe Abschnitt 7.6.2.3).
Der Fall, daß die Zuordnung der für die Ausbildung der Funktion notwendigen Aminosäuren von A auf Positionen von B nicht eindeutig möglich ist, kann aus folgenden Gründen eintreten:

1. Es sind zwar die für die Ausbildung der Funktion notwendigen Aminosäuretypen (zum Beispiel Serin, Histidin und Asparaginsäure in Serinproteasen) bekannt, aber deren genaue Position in der Sequenz A ist unbekannt.
2. Die Funktion ist nicht von bestimmten Aminosäuretypen abhängig und der oder die für die Funktion notwendigen Sequenzabschnitte können nur durch reguläre Ausdrücke beschrieben werden, die in der Sequenz von B an mehreren Stellen passen.

Dies sind Spezialfälle dessen, was für die in Datenbanken abgelegten Sequenzmuster als Regelfall anzusehen ist. Abschnitt 7.2.4.2 beschreibt, wie diese Information trotz vorhandener Mehrdeutigkeiten genutzt werden kann.
Im Fall 2(a) ist entweder die aktive Stelle durch ein nicht hinreichend eindeutiges Sequenzmuster beschrieben, oder die spezielle Funktion ist für das Protein B bisher unbekannt, beziehungsweise noch nicht annotiert worden. Falls es sich um ein sehr signifikantes Sequenzmuster handelt, das eine gewisse Länge aufweist und in dem viele Aminosäuretypen eindeutig bestimmt sind, wird analog zu Fall 1(a) verfahren. Handelt es sich um eine weniger signifikante Zuordnung oder paßt das Muster an verschiedenen Stellen in B, wird diese Information wie in Abschnitt 7.2.4.2 beschrieben genutzt.
Im Fall 2(b) kann die Information über die aktive Stelle von der RDP–Methode nur als Randbedingung für die Auswahl zulässiger Teillösungen verwendet werden (siehe Abschnitt 7.6.2.3).

7.2.4.2 Orakel und Information aus Datenbanken

Datenbanken wie `PROSITE` [17, 20] verbinden das Wissen über die Funktion oder Teilfunktionen eines Proteins mit bestimmten Mustern in der Sequenz. Diese Sequenzmuster werden als reguläre Ausdrücke über dem Alphabet der Aminosäuren beschrieben. Das Instrument der regulären Ausdrücke erlaubt in diesen Mustern für eine Position neben eindeutigen Aminosäuren, Mengen von Aminosäuren oder einfach nur Platzhalter für beliebige Aminosäuren zu spezifizieren. Zum Beispiel ist eine `ATP`–Bindestelle nach `PROSITE` in vielen Fällen durch den regulären Ausdruck `[AG]-x(4)-G-K-[STG]` charakterisiert (zur Erklärung: an erster Position

befindet sich ein Alanin oder Glycin, gefolgt von vier beliebigen Aminosäuren, gefolgt von einem Glycin, einem Lysin und entweder einem Serin, Threonin oder Glycin).
Mit Hilfe dieser regulären Ausdrücke werden im ersten Schritt alle möglichen Bindungsstellen in der bearbeiteten Sequenz gesucht. Diese Voranalyse kann effizient in Zeit $O(k * m * n)$ erfolgen [68], wobei k die Anzahl der regulären Ausdrücke in der Datenbank, m deren maximale Länge und n die Länge der Sequenz ist. Da die Muster mit Informationen zu ihrer Funktion annotiert sind, liefert diese Suche nicht nur die Positionen der möglichen Bindungsstellen in der Sequenz, sondern möglicherweise auch erste Hinweise auf die Funktion des untersuchten Proteins.
Im nächsten Schritt werden zu den in der Sequenz gefundenen möglichen Bindungsstellen potentielle Partner in der Sequenz der Struktur gesucht. Da viele der in der `PROSITE`–Datenbank gespeicherten regulären Ausdrücke nicht sehr spezifisch sind, werden Instanzen in der Regel häufiger in einer Sequenz gefunden. Daher kann über die meisten dieser möglichen Bindungsstellen auch keine eindeutige Zuordnung von Teilen der Sequenz auf Teile der Struktur festgelegt werden. Vielmehr ergeben sich verschiedene Kombinationsmöglichkeiten, um gleiche Bindungsstellen aufeinander abzubilden. Auf der Strukturseite kann die Auswahl möglicherweise eingeschränkt werden, da eine Bindungstelle immer von außerhalb des Proteins zugänglich sein sollte.
Eine Einschränkung auf die signifikanteste, in beiden Sequenzen gefundene Bindungstelle ist nicht sinnvoll, da viele Proteine, insbesondere Enzyme, nicht nur eine, sondern mehrere Bindungstellen zur Bindung verschiedenartiger Liganden enthalten. Da verschiedene mögliche Bindungsstellen für die Funktion eines Proteins von unterschiedlicher Bedeutung sein können, ist beim Vergleich verschiedener aber in Teilen funktionsverwandter Proteine darauf zu achten, daß zunächst die wichtigen Bindungsstellen aufeinander abgebildet werden.
Daher werden alle in der Sequenz gefundenen potentiellen Bindungsstellen pb_i in einer Liste PB_A gesammelt. Die Wahrscheinlichkeit $p\,(pb_i, A)$, das Aminosäuremuster der potentiellen Bindungsstelle pb_i in der Sequenz A zu finden, ist:

$$p\,(pb_i, A) = 1 - \left(1 - \prod_{j=1}^{|pb_i|} \sum_{a \in pb_{ij}} p(a)\right)^{|A|-|pb_i|+1}$$

Dabei ist $p(a)$ die Wahrscheinlichkeit einer Aminosäure $a \in \Sigma$ in natürlichen Proteinsequenzen (entlehnt aus `BLAST` [10]). Eine potentielle Bindungsstelle pb_i wird also durch die Anfangs– und Endposition, die Wahrscheinlichkeit $p\,(pb_i, A)$ ihres Auftretens in Sequenz A und die eindeutige `PROSITE`–Nummer charakterisiert.
Die in der Sequenz der Struktur B gefundenen äquivalenten potentiellen Bindungsstellen werden den entsprechenden Listeneinträgen von PB_A zugeordnet. Die Wahrscheinlichkeiten, ein bestimmtes Bindungsstellenmuster in Sequenz A oder B zu finden, werden als unabhängig angenommen, so daß sich die Wahrscheinlichkeit $p\,(pb_i, A, B)$, Bindungsstellenmuster pb_i in den Sequenzen A und B

zu finden, als Produkt der Wahrscheinlichkeiten $p(pb_i, A)$ und $p(pb_i, B)$ ergibt. Die RDP–Methode verwendet die so gefundenen Paare aus PB_A auf folgende Arten:

- Die möglichen Zuordnungen, deren Wahrscheinlichkeit einen Grenzwert unterschreiten, werden wie Ergebnisse anderer Orakel verwendet.
- Zunächst wird eine „maximal"–mögliche Zuordnung von potentiellen Bindungstellen berechnet (siehe Abschnitt 7.3.5), wobei ihre Wahrscheinlichkeiten und die durch ihre Zuordnung hervorgerufenen Verschiebungen zwischen Sequenz und Struktur in die Kostenfunktion eingehen. Die gefundene Zuordnung wird wie ein Ergebnis anderer Orakel verwendet.

7.3 Vorselektion der Teillösungen

In Abschnitt 7.2 wurde eine Vielzahl von Orakeln beschrieben, die unterschiedliche partielle Lösungen für ein Teilproblem generieren. Jede dieser partiellen Lösungen $PA_v^\wedge$ entspricht einem Knoten $v \in V_\wedge$ des Lösungsbaums $\mathcal{T}$, der wiederum Wurzel eines ganzen Teilbaums sein kann. Die Größe des Teils des Lösungsbaums, der explizit aufgebaut wird, bestimmt wesentlich die Laufzeit. Daher kommt es darauf an, diesen frühzeitig einzuschränken, indem

- nicht signifikante,
- identische,
- nicht zulässige und
- sehr ähnliche Teillösungen

direkt erkannt und entsprechend behandelt werden. Im folgenden werden die von der RDP–Methode zu diesem Zweck verwendeten Kriterien und Methoden beschrieben. Wenn dabei davon die Rede ist, daß eine Teillösung verworfen oder gelöscht wird, so bedeutet das, daß der zugehörige Knoten $v \in V_\wedge$ aus dem Lösungsbaum $\mathcal{T}$ entfernt wird und so auch der zugehörige Teilbaum des Lösungsbaums nicht erzeugt wird. In Abschnitt 7.3.5 wird zusätzlich eine allgemeine Methode beschrieben, wie – alternativ zum rekursiven Abstieg – unterschiedliche Teillösungen zu einer umfangreicheren Lösung kombiniert werden können.

7.3.1 Signifikanz von Teillösungen

Einigen Orakeln (zum Beispiel denen, die aus der Information über potentielle Bindungstellen abgeleitet sind) kann direkt durch die Spezifizität des Sequenzmusters eine Signifikanz in Abhängigkeit von der Wahrscheinlichkeit, dieses Muster zufällig zu beobachten, zugeordnet werden. In diesen Fällen wird die Entscheidung über die weitere Verwendung eines Orakelergebnisses anhand des Vergleichs der Signifikanz gegen einen vorgegebenen Mindestwert getroffen.

Schwieriger wird die Entscheidung, wenn die partielle Zuordnung von einem der anderen Orakeltypen erzeugt wurde, da dann nicht so einfach mit dem Wahrscheinlichkeitsargument gearbeitet werden kann. Eine Bewertung der Signifikanz einer Lösung durch Vergleich gegen die Wahrscheinlichkeit, diese Zuordnung zufällig zu finden, ist für die typischen, mit Sequenzstrukturalignmentmethoden bearbeiteten Problemstellungen nur schwer möglich, da hohe Sequenzidentitäten unwahrscheinlich sind, und die Statistik aller anderen Kriterien bislang nur unzureichend untersucht ist. Daher müssen, wenn die Signifikanz einer Teillösung sich nicht direkt aus dem Orakel ergibt, wie dies zum Beispiel bei der Beschränkung auf die zuverlässigen Bereiche eines Alignments der Fall ist, andere Wege gefunden werden.
Die besondere Eigenschaft partieller oder lokaler Lösungen eines Problems ist, daß der positionelle Score der lokalen Lösung signifikant höher als der positionelle Score einer globalen Lösung des Teilproblems ist. Ist dies nicht der Fall, so existieren in dem bearbeiteten Teilproblem keine lokalen Bereiche, die aufgrund der bisher verfügbaren Information mit höherer Zuverlässigkeit zugeordnet werden können, als dies unter Einbeziehung der ganzen das Teilproblem definierenden Sequenz– und Strukturbereiche möglich ist.
Zur Abschätzung der Signifikanz wird daher von der RDP–Methode neben der partiellen Lösung auch der positionelle Score der globalen Lösung eines Teilproblems berechnet. Ist der positionelle Score der partiellen Lösung signifikant höher als der der globalen Lösung, so wird die partielle Lösung als signifikant eingestuft, das aktuelle Teilproblem wird anhand der gefundenen Teillösung in Unterprobleme zerlegt und diese werden in der RDP–Methode weiterbearbeitet.
Ein nicht signifikanter Unterschied zwischen dem positionellen Score der lokalen und dem der globalen Lösung des Teilproblems kann zwei Gründe haben:

1. Es ist keine eindeutige Zuordnung der Teilbereiche anhand der für das Orakel verwendeten Informationen möglich.
2. Die partielle Lösung entspricht nahezu einer vollständigen Zuordnung, das heißt nur wenige Positionen der Sequenz beziehungsweise der Struktur des Teilproblems sind nicht Bestandteil der lokalen Lösung.

Im ersten Fall werden sowohl die partielle als auch die globale Lösung verworfen. Falls keines der anderen Orakel eine sinnhafte Lösung des Teilproblems generiert, wird auch das Teilproblem aus dem Lösungbaum entfernt. Falls mit anderen Orakel signifikante Teillösungen gefunden werden, so werden nur diese weiterverfolgt.
Im zweiten Fall ist eine weitere Untersuchung der resultierenden Unterprobleme, die nur aus kurzen Abschnitten der Sequenz beziehungsweise der Struktur bestehen, wenig sinnvoll und wird nicht durchgeführt. Daher wird entweder die partielle Lösung zu einer globalen Lösung erweitert, indem die nicht alinierten Bereiche als Insertionen beziehungsweise Deletionen markiert werden und so als nicht zugeordnete Bereiche zurückbleiben, oder die globale Lösung wird als eine mögliche Lösung des Teilproblems übernommen.

7.3.2 Identische Teillösungen

Obwohl in den Orakeln zur Bestimmung der partiellen Lösungen für ein Teilproblem unterschiedliche Kostenfunktionsbestandteile verwendet werden, kommt es dennoch vor, daß unterschiedliche Orakel gleiche Lösungen generieren. Zur effizienten Erkennung gleicher Lösungen wird jeder Lösung ein *Hashcode* zugeordnet, der sich aus der gewichteten Summe ihrer Anfangs– und Endpositionen und der die Lösungen kodierenden Indizes errechnet. Stimmt dieser *Hashcode* für zwei Teillösungen überein, wird überprüft, ob die Teillösungen identisch sind.
Kodieren zwei Söhne eines Knotens $p \in V_V$ für identische Teillösungen, so wird einer von ihnen direkt aus $\mathcal{T}$ gelöscht, bevor mit der Erzeugung des durch ihn aufgespannten Teilbaums begonnen wird. Daß unterschiedliche Orakel die gleiche Lösung liefern, sagt aber auch etwas über die Signifikanz dieser Lösung. Daher erhalten die aus dieser Lösung resultierenden Unterprobleme einen *Bonus*:

- Die Anzahl der maximal von den Orakel zur Lösung der Unterprobleme generierten partiellen Lösungen wird um 50% erhöht.
- Die Teilprobleme werden durch Multiplikation des Bonusfaktors mit 1.5 in der Prioritätsliste der offenen Teilprobleme vorgezogen und so vorrangig bearbeitet (siehe Abschnitt 7.5).
- Bei dem später beschriebenen Auswahlschritt in der *bottom up*–Phase des Algorithmus kann eine ebenfalls um 1.5 erhöhte Anzahl von Lösungen an die nächst höhere Hierarchiestufe im Lösungsbaum weitergereicht werden (siehe Abschnitt 7.6.2).

7.3.3 Nicht zulässige Teillösungen

Die Erkennung und Aussonderung nicht zulässiger Teillösungen ist auch ein wesentlicher Bestandteil der *bottom–up*–Phase der RDP–Methode. Sowohl die Zulässigkeitskriterien als auch die zur ihrer Auswertung verwendeten Verfahren werden daher in Abschnitt 7.6.2 beschrieben. In der *top–down*–Phase der RDP–Methode werden nach diesen Kriterien nicht zulässige Orakellösungen direkt verworfen.

7.3.4 Ähnliche Teillösungen

In der Regel werden in der RDP–Methode alle partiellen Lösungen, die von Orakeln generiert werden und die in den vorangegangenen Abschnitten beschriebenen Filter überstanden haben, auch weiterverfolgt. Die Einschränkung der Alternativen und damit der Größe des Lösungsbaums erfolgt bereits bei der Generierung der Lösungen durch die Auswahl der verwendeten Orakel und die Einschränkung der für jedes Orakel erlaubten Lösungen.
Optional ist es aber auch möglich, die Anzahl der Teillösungen vor dem rekursiven Abstieg einzuschränken, indem ähnliche Teillösungen zusammengefaßt werden. Da die Anzahl der unterschiedlichen und zulässigen Teillösungen für ein

Teilproblem in der typischerweise verwendeten Parametrisierung der Methode zwischen 5 und 25 beträgt, ist es möglich, den vollständigen paarweisen Vergleich der Teillösungen durchzuführen.
Als Ähnlichkeitsmaß wird die Anzahl der in je zwei Teillösungen identisch zugeordneten Strukturpositionen im Verhältnis zur minimalen Anzahl der in einer der beiden Teillösungen zugeordneten Positionen verwendet. Dieses Ähnlichkeitsmaß kann für zwei Teillösungen durch einfachen Vergleich der Indexvektoren, die die jeweilige Teillösung kodieren, berechnet werden.
Da jeweils nur sehr ähnliche Teillösungen zu einer Teillösung zusammengefaßt werden, kann dies in einer einfachen, dem *single linkage clustering* [178] ähnlichen Vorgehensweise erfolgen:

1. Bestimme die zwei ähnlichsten Teillösungen.
2. Falls die zwei Teillösungen unähnlicher als ein vorgegebener Grenzwert (zum Beispiel 80%) sind, beende die Zusammenfassung der Teillösungen.
3. Falls sie ähnlicher sind, verfahre wie folgt:
   (a) Falls sich die Anzahl der zugeordneten Positionen in beiden Teillösungen um mehr als die Hälfte unterscheidet, wird die Ähnlichkeit dieser Teillösungen für die nachfolgenden Schritte zu 0 gesetzt und beide Teillösungen werden beibehalten, da die Teile der größeren Teillösung nur von einem der Orakel zugeordnet wurden und das andere Orakel eine Unterteilung in weitere Unterprobleme nahelegt.
   (b) Sonst werden die gemeinsamen Zuordnungen der beiden Teillösungen in eine neue Teillösung übernommen, die beiden Teillösungen gelöscht und die Ähnlichkeit der neuen Teillösung zu den anderen Teillösungen bestimmt.
4. Beginne bei Punkt 1.

Werden zwei ähnliche Teillösungen zu einer Teillösung zusammengefaßt, wird den zugehörigen Unterproblemen wie auch beim Löschen gleicher Teillösungen ein Bonus gewährt (siehe Abschnitt 7.3.2).
In praktischen Anwendungen hat sich jedoch die Einschränkung der Teillösungen bereits bei der Generierung als sinnvoller erwiesen. Da die Anzahl der unterschiedlichen Orakel begrenzt ist, kann eine größere Vielfalt an Teillösungen nur durch die Einbeziehung suboptimaler Teillösungen erzeugt werden. Optimale und suboptimale Lösungen, die vom gleichen Orakel erzeugt wurden, sind jedoch in der Regel sehr ähnlich und werden bei einer Zusammenfassung, wie sie oben beschrieben wird, direkt wieder eliminiert.

7.3.5 Kombination verschiedener Orakellösungen

In diesem Abschnitt wird eine allgemeine Möglichkeit vorgestellt, wie neben dem rekursiven Abstieg für ein Teilproblem Lösungen bestimmt werden können, die in einem Schritt größere Bereiche einander zuordnen, als dies durch die von einem Orakel berechneten partiellen Lösung geschieht. Dazu werden partielle Lösungen, die unterschiedliche Bereiche in Sequenz und Struktur zuordnen, zu einer umfassenderen Lösung eines Problems zusammengefaßt.

Definition 7.8 (Kombinierbare partielle Lösungen)
Seien f und $g \in \mathcal{H}$ partielle Lösungen für ein Teilproblem SP bestehend aus einem Sequenzabschnitt A und einem Strukturausschnitt B.
Bezeichne $min_A(f) = min_{f(i)\neq\emptyset}\ i$ ($max_A(f) = max_{f(i)\neq\emptyset}\ i$) die erste (letzte) von f zugeordnete Position aus A und $min_B(f) = min_{f^{-1}(j)\neq\emptyset}\ j$ ($max_B(f) = max_{f^{-1}(j)\neq\emptyset}\ j$) die erste (letzte) von f zugeordnete Position aus B. O.B.d.A. sei $min_A(f) < min_A(g)$.
Zwei partielle Lösungen f und g heißen kombinierbar *genau dann, wenn eine der folgenden Bedingungen erfüllt ist:*

$$(i) \quad max_A(f) \leq min_A(g) \ \wedge \ max_B(f) \leq min_B(g)$$
$$(ii) \quad \exists\, i \in A \ \ \forall\, j \in A \wedge j \geq i \ : f(j) = g(j) \ \vee \ f(j) = \emptyset \ \vee \ g(j) = \emptyset \ .$$

$OV(f,g)$ bezeichnet die Anzahl der überlappenden Positionen von f und g und $NG(f,g)$ die Anzahl der aus der Kombination von f und g resultierenden Insertionen beziehungsweise Deletionen. $NG(0,f)$ ($NG(f,\infty)$) ist die Anzahl der Insertionen beziehungsweise Deletionen, die bei Verwendung von f als erster (letzter) Teillösung für SP vor (nach) f notwendig werden.

Nach Definition 7.8 sind zwei partielle Lösungen kombinierbar, wenn sie entweder sowohl auf Sequenz– als auch Strukturseite disjunkt sind (Bedingung (i)) oder ab einer bestimmten Position die Zuordnungen identisch sind oder nur in einer der Lösungen existieren (Bedingung (ii)). Sind zwei Lösungen nach Bedingung (i) kombinierbar, so ist gleichzeitig durch die sequentielle Abfolge eine Ordnung auf diesen Lösungen definiert. Durch die lexikographische Ordnung auf den jeweiligen Anfangs– und Endpunkten kann diese Ordnung einfach auf nach Bedingung (ii) kombinierbare Lösungen erweitert werden.
Die Kombination zweier partieller Lösungen f und g kann bei Erhaltung der durch die Peptidkette vorgegebenen Topologie bewirken, daß $NG(f,g)$ sehr groß wird, was bedeutet, daß große Teile der Sequenz oder der Struktur nicht mehr zuzuordnen sind. Bei der Bestimmung einer „maximalen“ Menge kombinierbarer partieller Lösungen ist daher nicht nur auf die Gewichte der Lösungen, sondern auch auf die Anzahl der aus der Zusammenfassung resultierenden unvermeidlichen Insertionen und Deletionen zu achten.

Das Gewicht $\omega(f)$ einer Lösung f hängt vom jeweils bearbeiteten Problem ab und ist häufig nur schwer festzulegen. So mißt ein Biologe, der an einem bestimmten Protein interessiert ist und schon einiges über seine mögliche Funktion weiß, der Abbildung verschiedener Bindungsstellen pb_i unterschiedliche Bedeutung bei. In der Regel verwendet die RDP–Methode hier jedoch den negativen Logarithmus des *p–values* (siehe Abschnitt 7.2.1.1) oder der Wahrscheinlichkeit, eine potentielle Bindungsstelle in beiden Proteinen zu finden (siehe Abschnitt 7.2.4.2), als Gewicht $\omega(f)$ einer partiellen Zuordnung f.
Im folgenden wird eine allgemeine Methode beschrieben, wie diese oder andere Gewichte $\omega(f)$ bei der Berechnung einer „maximalen“ Menge kombinierbarer Lösungen berücksichtigt werden können. Dazu wird das Auswahlproblem als gerichteter Graph $G_C = (V_C, E_C, \omega, \nu)$ beschrieben, in dem die Knoten eines Weges $W(s \to t)$ zwischen den zusätzlichen Start– und Zielknoten s und t, eine Menge von kombinierbaren partiellen Lösungen repräsentieren. Die Funktion $\omega(f)$ belohnt die Verwendung einer Lösung f, und die Funktion $\nu(f, g)$ bestraft die für die gleichzeitige Verwendung zweier Lösungen f und g notwendigen Insertionen und Deletionen.
Das Problem der Berechnung der maximalen Kombination partieller Lösungen kann damit als *Kürzeste–Wege*–Problem beschrieben werden:

Definition 7.9 (Maximale Kombination partieller Lösungen)

Gegeben: *Ein gerichteter Graph* $G_C = (V_C, E_C, \omega, \nu)$ *mit:*

- $V_C = \{s, t\} \cup \{f \mid f \text{ Teillösung}\}$
- $E_C = \{(s, f) \mid f \text{ Teillösung}\} \cup \{(f, t) \mid f \text{ Teillösung}\} \cup \{(f, g) \mid f, g \text{ kombinierbare Teillösungen}\}$
- *Knotengewichten* $\omega(u) = \begin{cases} 0 & : \quad u = s \vee u = t \\ \omega(u) & : \quad u \text{ Teillösung} \end{cases}$
- *Kantengewichten*
$$\nu(u, v) = \begin{cases} NG(0, v) & : \quad u = s \wedge v \text{ Teillösung} \\ NG(u, \infty) & : \quad v = t \wedge u \text{ Teillösung} \\ NG(u, v) - OV(u, v)) & : \quad u, v \text{ Teillösungen} \end{cases}$$

Gesucht: *Ein kürzester Weg* $KW(s \to t) \subseteq G_C$ *von* s *nach* t*, wobei die Länge eines Weges* $W \subseteq G_C$ *durch* $\sum_{u,v \in W} \nu(u, v) - \lambda \sum_{u \in W} \omega(u)$ *bestimmt ist.*

Die Knotengewichte ω gehen mit negativem Faktor λ gewichtet gegen die Kantenkosten in die Kosten eines Weges von s nach t ein. Über den Gewichtungsfaktor λ werden die Gewichte der Lösungen gegen die notwendigen Insertionen und Deletionen gewichtet und so festgelegt, welche Kosten für Insertionen und Deletionen für einen Gewinn durch die Zusammenfassung zweier Teillösungen in Kauf genommen werden.

Die Kantengewichte ν eines Weges im Graphen G_C modellieren lineare Gapkosten für die Anzahl der Insertionen und Deletionen, die durch die Kombination von partiellen Lösungen notwendig werden. Bei der Bestimmung des kürzesten Weges

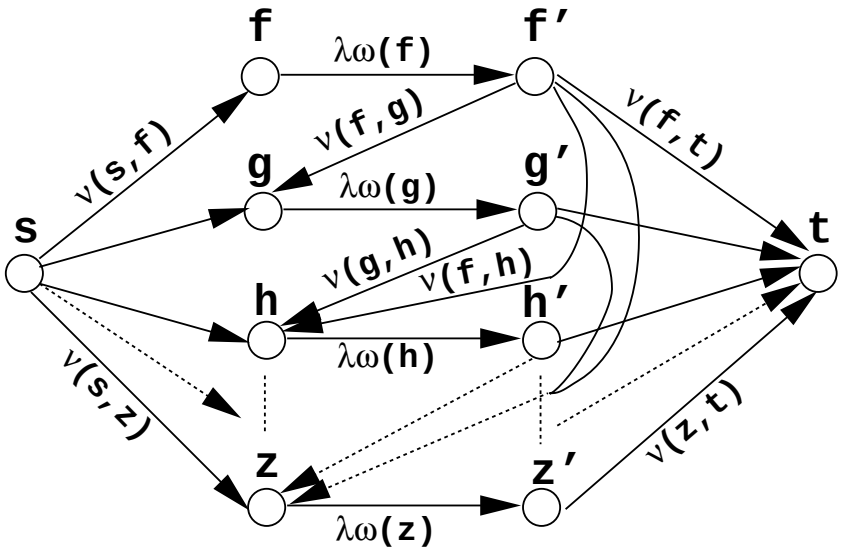


Abbildung 7.9: Zusammenfassung von Teillösungen für ein Teilproblem als *kürzeste–Wege*–Problem.

werden die Knotengewichte in Kantengewichte transformiert, indem ein Knoten f zu zwei Knoten f und f' expandiert wird und die Kante (f, f') das Kantengewicht $\lambda\omega(f)$ erhält (siehe Abbildung 7.9).
Die Kanten sind gerichtet, da beim Sequenzstrukturalignment in der Regel die Reihenfolge der Reste in Sequenz und Struktur erhalten bleiben soll und sich diese Ordnungsrelation auf die Teillösungen überträgt. Da G_C ein gerichteter azyklischer Graph ist, kann der kürzeste Weg von s nach t in Laufzeit $\Theta(|V_C| + |E_C|)$ bestimmt werden [68]. Für typische Anwendungsbeispiele der RDP–Methode ist die Anzahl der unterschiedlichen Teillösungen für ein Teilproblem kleiner als 25. Da außerdem nicht alle Teillösungen kombinierbar sind, ist die Anzahl der Kanten des Graphen in der Regel kleiner als 200.
Im einfachsten Fall enthält der kürzeste Weg genau eine partielle Lösung. Dann ändert sich an den bisher für das Teilproblem *SP* gefundenen Lösungen nichts. Enthält der gefundene kürzeste Weg mehr als eine Lösung, so werden die enthaltenen Lösungen zu einer Lösung kombiniert (siehe dazu auch Abschnitt 7.6.1) und aus der Liste der alternativen Lösungen eines Teilproblems gelöscht. So entstandene partielle Lösungen enthalten in der Regel Wechselgaps, das heißt Bereiche aus Sequenz und Struktur, für die noch keine Zuordnung getroffen werden konnte. Bei der Aufteilung in noch zu lösende Unterprobleme (siehe Abschnitt 7.4) werden durch diese Wechselgaps neben den durch die Bereiche rechts und links der partiellen Lösung definierten Unterproblemen zusätzliche neue Unterproble-

me definiert, so daß der Knoten des Lösungsbaums, der die partielle Lösung repräsentiert, eventuell mehr als zwei Unterprobleme hat.
Wie oben beschrieben, kann jedoch nicht nur mit dem kürzesten Weg, sondern analog auch mit den suboptimalen Wegen verfahren werden, so daß die Kombination von Orakellösungen wie alle anderen Orakel zu einer Menge möglicher Teillösungen führt mit dem Unterschied, daß diese mehr Positionen in Sequenz und Struktur in die Teillösung einbeziehen.

7.4 Aufteilung in Unterprobleme

Die Funktion $g_\vee$ wird, wie bereits bei der Beschreibung des Lösungsbaums erwähnt (siehe Abschnitt 7.1.5), auf Knoten $v \in V_\wedge$ angewendet und zerlegt ein Teilproblem $SP_u^\vee$ mit $(u, v) \in E^{down}_{\vee\to\wedge}$ anhand der partiellen Zuordnung $PA_v^\wedge$ in m Unterprobleme.
Daher ist an die partielle Zuordnung $PA_v^\wedge$ die Minimalforderung zu stellen, daß durch sie mindestens ein Paar von Positionen in Sequenz und Struktur identifiziert wird, das eine Aufteilung in kleinere Unterprobleme definiert, die rechts beziehungsweise links der durch das Paar definierten Schnittstelle liegen. Diese Bedingung ist auch dann erfüllt, wenn die partielle Zuordnung keine echte Zuordnung von Positionen ist, sondern anderweitig eine Schnittstelle definiert, zum Beispiel durch Festlegung von Domänengrenzen in Sequenz und Struktur.

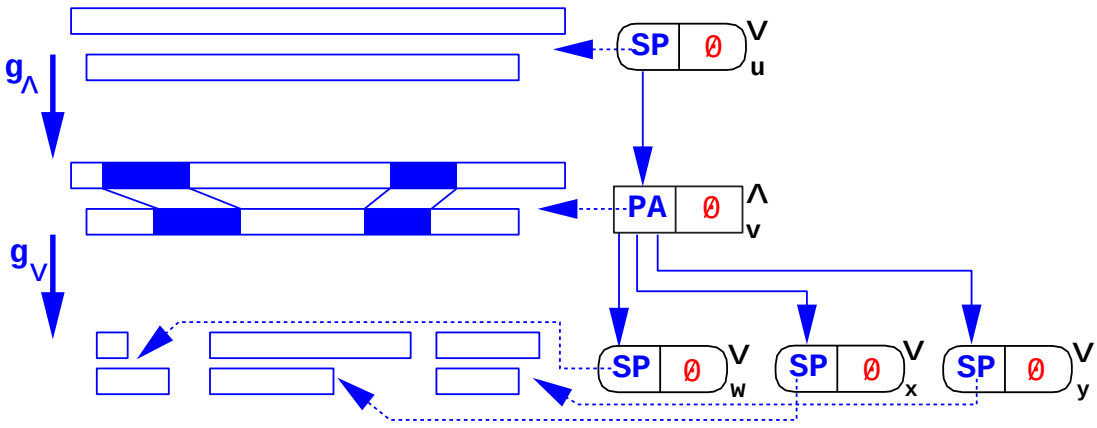


Abbildung 7.10: Aufteilung eines Teilproblems $SP_u^\vee$ anhand einer partiellen Lösung $PA_v^\wedge$ in Unterprobleme $SP_w^\vee, \ldots, SP_y^\vee$ (gefüllte Rechtecke kennzeichnen durch $PA_v^\wedge$ zugeordnete Bereiche).

Abbildung 7.10 zeigt diese durch die Funktion $g_\vee$ durchgeführte Aufteilung an einer partiellen Lösung $PA_v^\wedge$. Falls $PA_v^\wedge$ Struktur– und Sequenzpositionen aus $SP_u^\vee$ einander zuordnet, werden diese in den Teilsequenzen und –strukturen aus $SP_u^\vee$ als gelöst markiert, und die verbleibenden zusammenhängenden Teilstücke definieren dann die neuen Unterprobleme $SP_s^\vee$ (siehe Abschnitt 7.1.3). In der Regel soll die

Topologie der Proteinketten erhalten bleiben. Daher ist die Zuordnung der ein Teilproblem definierenden Teile der Sequenz und der Struktur eindeutig, und jedes Teilstück gehört zu genau einem Unterproblem.
Falls $PA_v^{\wedge}$ nur Schnittstellen in Sequenz und Struktur des Teilproblems $SP_u^{\vee}$ identifiziert, erfolgt die Aufteilung in Unterprobleme an diesen Schnittstellen.
Die Unterprobleme $s = (SP_s^{\vee}, \emptyset)$ werden als Söhne von dem Knoten v in den Lösungsbaum $\mathcal{T}$ und in die Prioritätsliste der offenen Teilprobleme (siehe dazu auch Abschnitt 7.5) eingetragen, falls sowohl die Teilsequenz als auch die Teilstruktur des jeweiligen Unterproblems eine gewisse Mindestlänge aufweisen, die noch signifikante Teillösungen für das Teilproblem möglich erscheinen läßt. Teilprobleme, bei denen zum Beispiel entweder die Teilsequenz oder die Teilstruktur kürzer als drei Aminosäurereste ist, werden daher direkt verworfen und somit weder in $\mathcal{T}$ noch in die Prioritätsliste der offenen Teilprobleme eingetragen. Bei ihrer Entstehung erben die Söhne von ihrem Vaterknoten sowohl den Bonusfaktor als auch die Anzahlen der maximal pro Orakeltyp erlaubten partiellen Zuordnungen.

7.5 Abarbeitungsreihenfolge des Lösungsbaums

Der Lösungsbaum $\mathcal{T}$ enthält zwei Arten von Knoten (siehe Abschnitt 7.1.5). Die *Und*–Knoten der Knotenmenge $V_{\wedge}$ enthalten in der *top–down*–Phase partielle Lösungen, anhand derer die zugehörigen Teilprobleme in Unterprobleme aufgeteilt werden. Die *Oder*–Knoten der Knotenmenge $V_{\vee}$ repräsentieren diese Teilprobleme und Unterprobleme.
In der gegenwärtigen Implementierung der RDP–Methode werden die *Oder*–Knoten aus den *Und*–Knoten erzeugt, sobald die partielle Zuordnung berechnet ist, da die Aufteilung in Unterprobleme für das Sequenzstrukturalignmentproblem unabhängig vom Rest des Lösungsbaums ist. Der Zustand des Lösungbaums kann jedoch über die Kontaktrelationen sehr wohl Einfluß auf die in den Orakeln berechneten partiellen Lösungen haben (siehe Abschnitt 7.2.3). Daher werden im folgenden mögliche Abarbeitungsreihenfolgen der offenen Teilprobleme diskutiert. Offene Teilprobleme entsprechen dabei den Knoten aus $V_{\vee}$, die zum betrachteten Zeitpunkt Blätter des Lösungsbaums sind.
Für die Abarbeitung der offenen Teilprobleme kennt die RDP–Methode zwei grundsätzlich verschiedene Modi. Im ersten Abarbeitungsmodus werden die offenen Teilprobleme in einer Prioritätsliste *pq* verwaltet. Im einfachsten Fall werden Unterprobleme dabei in der Reihenfolge bearbeitet, in der sie durch die Aufteilung in Unterprobleme (siehe Abschnitt 7.4) erzeugt werden. In diesem Fall wird der Lösungsbaum in Breitensuche (`BFS`) durchlaufen. Bei der Breitensuche wird mit der Zusammenfassung von Lösungen und damit der Reduktion des benötigten Speicherplatzes erst nach Abarbeitung der Liste der offenen Probleme begonnen. Daraus resultiert das Problem, daß der Lösungsbaum insbesondere für Sequenzstrukturpaare, die sich als nicht verwandt herausstellen, sehr schnell sehr groß

wird. Dies liegt daran, daß die gefundenen partiellen Lösungen für nicht verwandte Proteine in der Regel sehr kurz sind, somit die Problemgröße nur langsam abnimmt und damit der Lösungsbaum zwischenzeitlich sehr groß wird.
Daher werden von der RDP–Methode nicht alle Teilprobleme in Breitensuche abgearbeitet, sondern nach einer vorgegebenen Anzahl von Teilproblemen (standardmäßig 500) wird dazu übergegangen, alle nach diesem Zeitpunkt erzeugten Unterprobleme unter Reduktion der vom Vaterknoten ererbten Anzahl erlaubter alternativer Orakel und im zweiten Abarbeitungsmodus — der Tiefensuche (`DFS`) — zu lösen. Die zu diesem Zeitpunkt in der Prioritätsliste enthaltenen Teilprobleme (ein Vielfaches von 500) sind von der Änderung des Abarbeitungsmodus nicht betroffen. Die Anzahl der Prioritätslistenelemente ist ein Vielfaches der bereits abgearbeiteten Teilprobleme, da jede bei der Abarbeitung eines Teilproblems berechnete Teillösung in der Regel zwei neue Unterprobleme zur Folge hat.
Im Tiefensuchemodus werden Teilbäume des Lösungsbaums, für die bereits alle Unterprobleme gelöst sind, sofort zu Gesamtlösungen für die jeweilige Wurzel dieses Teilbaums zusammengefaßt, wie dies für den Rest des Lösungsbaums erst in der *bottom–up*–Phase geschieht (siehe Abschnitt 7.6).
Die Anzahl der untersuchten partiellen Lösungen wird außerdem für Teilprobleme reduziert, bei denen entweder die Struktur oder Sequenz sehr kurz ist, zum Beispiel kürzer als die doppelte Mindestlänge für offene Teilprobleme ist (siehe Abschnitt 7.4).
Die Verwaltung der offenen Teilprobleme in einer Prioritätsliste bietet den weiteren Vorteil, daß die Teilprobleme nicht nur in der Reihenfolge ihrer Entstehung abgearbeitet werden können, sondern auch andere Kriterien zur Festlegung der Abarbeitungsreihenfolge verwendet werden können:

- Ein erstes Kriterium ist das Minimum der Längen von Teilstruktur und Teilsequenz. Die Idee dabei ist, daß für größere Teilprobleme eher noch gute Teillösungen zu erwarten sind.
- Ein alternatives Sortierkriterium ist die Anzahl der für die Strukturpositionen eines Teilproblems SP bereits eindeutig festgelegten Kontaktpartner (also die Mächtigkeit der Menge B^1_{SP}). Je mehr externe Kontaktpartner bekannt sind, desto gezielter kann die Zuordnung von Teilen der Teilsequenz zu Strukturpositionen des Teilproblems bezüglich des Kontaktpotentials erfolgen.

Während das erste Kriterium einfach zu berechnen ist und auch für ein Teilproblem SP immer konstant bleibt, erfordert das zweite Kriterium die Bestimmung der Menge B^1_{SP} (siehe Abschnitt 7.2.3.1). Außerdem ändert sich die Mächtigkeit von B^1_{SP} dynamisch während der Abarbeitung des Lösungsbaums, so daß nicht nur die Prioritätsliste immer neu sortiert werden muß, sondern gegebenenfalls auch das Sortierkriterium neu auszuwerten ist.
Durch multiplikative Gewichtung des ausgewählten Sortierkriteriums mit dem bereits erwähnten *Bonusfaktor* kann die Abarbeitungsreihenfolge der Teilprobleme

von anderen Unterfunktion der RDP–Methode beeinflußt, die zum Beispiel identische Teillösungen entfernen (siehe Abschnitt 7.3.2) oder sehr ähnliche Teillösungen zusammenfassen (siehe Abschnitt 7.3.4).
Bei den bisher durchgeführten Experimenten lieferte das einfache Längenkriterium etwas bessere Ergebnisse, wenn man die Güte des berechneten Sequenzstrukturalignments nicht anhand des von der Alignmentmethode optimierten Bewertungskriteriums, sondern anhand der zugeordneten Reste und der *RMS*–Abweichung der Superposition der Strukturen (vorausgesetzt diese ist auch für die Sequenz bekannt) mißt. Dies ist ein weiteres Indiz dafür, daß die alleinige Optimierung der Paarwechselwirkungen nicht gleichbedeutend mit der Berechnung von Lösungen ist, die eine niedrige *RMS*–Abweichung aufweisen und als biologisch sinnvoll angesehen werden können.

7.6 Bestimmung der Gesamtlösung

Die *bottom–up*–Phase der RDP–Methode kombiniert beginnend an den Blättern des Lösungsbaums $\mathcal{T}$ Teillösungen zu größeren Lösungen und baut dabei den Lösungsbaums $\mathcal{T}$ rekursiv ab. Die *bottom–up*–Phase schließt mit einem oder mehreren Lösungsvorschlägen A für das bearbeitete Sequenzstrukturalignmentproblem P in der Wurzel (P, A) des Lösungsbaums $\mathcal{T}$ (siehe Abbildung 7.4).
Wie bereits in Abschnitt 7.1.5 beschrieben, besteht die *bottom–up*–Phase aus den Funktionen $f_\wedge$, $e_\wedge$, $f_\vee$ und $e_\vee$, wobei der Index der Funktion jeweils den Knotentyp des Knotens angibt, der das Ergebnis der Funktion repräsentiert. Für im `DFS`–Modus abgearbeitete Knoten aus $\mathcal{T}$ werden die Funktionen der *bottom–up*–Phase ausgeführt, sobald sie anwendbar sind (siehe Abschnitt 7.1.5). Die Funktionen der *bottom–up*–Phase sind auf einen Knoten anwendbar, sobald sie vollständig auf allen seinen Sohnknoten durchgeführt wurden.
Eine entgültige Paarpotentialbewertung einer Teillösung ist erst dann möglich, wenn die Belegung aller Kontaktpartner eindeutig festgelegt ist. Dies kann zwar bis zum Abschluß der Berechnung nicht garantiert werden, doch steigt die Anzahl der eindeutig alinierten Kontaktpartner mit dem Fortschreiten des Verfahrens. Daher wird die Auswertung der Knoten, die in der *top–down*–Phase im `BFS`–Abarbeitungsmodus erzeugt werden, solange wie möglich hinausgezögert und erfolgt in einem zusätzlichen `DFS`–Durchlauf nach Beendigung der *top–down*–Phase.
In der gegenwärtigen Realisierung der RDP–Methode berechnet die Funktion $f_\vee$ die Menge der Lösungen $TA_v^\vee$ für Teilproblem $SP_v^\vee$ durch Vereinigung der $TA_w^\wedge$ über alle Söhne w von v und die Funktion $e_\wedge$ ist die Identitätsfunktion. Daher werden die wesentlichen Schritte der *bottom–up*–Phase durch die Funktionen $f_\wedge$ und $e_\vee$ durchgeführt:

- $f_\wedge$ kombiniert die Lösungen $TA_w^\vee$ von Söhnen w von $v \in V_\wedge$ mit der partiellen Lösung $PA_v^\wedge$ zu Lösungen $TA_v^\wedge$. Abschnitt 7.6.1 beschreibt diese Kombination.

- $e_\vee$ schränkt die an Knoten $v \in V_\vee$ gespeicherte Menge von Lösungen $TA_v^\vee$ auf die zulässigen und besseren Lösungen ein. Die so bestimmte Auswahl von Lösungen ist Eingabe der Funktion $f_\vee$ auf der nächsthöheren Hierarchiestufe. Abschnitt 7.6.2 beschäftigt sich mit diesem Auswahlschritt.

7.6.1 Kombination von Lösungen an Knoten aus $V_\wedge$

Die Funktion $f_\wedge$ wird auch als *Merge*–Funktion (siehe Abschnitt 6.2) bezeichnet, da sie partielle Zuordnungen eines Teilproblems mit den Lösungen, der zugehörigen Unterprobleme kombiniert oder zusammenmischt (siehe auch Abbildung 6.6).

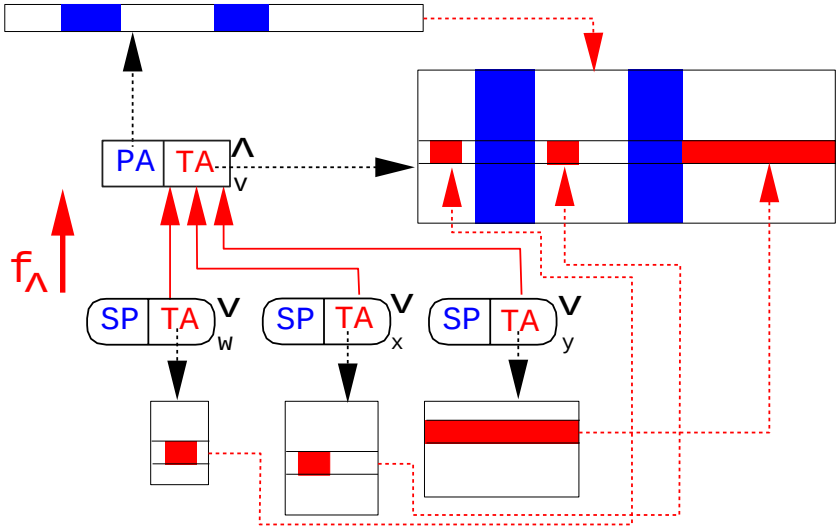


Abbildung 7.11: Bestimmung einer Lösung aus $TA_v^\wedge$ durch Kombination einer partiellen Lösung $PA_v^\wedge$ mit je einer Lösung aus $TA_w^\vee$, $TA_x^\vee$ und $TA_y^\vee$.

Wie Abbildung 7.11 andeutet, enthalten Unterprobleme $SP_w^\vee$ mit $(v,w) \in E_{\wedge\to\vee}^{down}$ — damit auch ihre Lösungen $TA_w^\vee$ — nach ihrer Definition nur Sequenz– und Strukturpositionen, für die durch die partielle Lösung $PA_v^\wedge$ keine Zuordnung gefunden wurde, das heißt die entsprechenden Einträge in $PA_v^\wedge$ sind noch leer (siehe Abschnitt 7.1.4). Daher kann eine Lösung aus $TA_v^\wedge$ durch Erstellen einer Kopie von $PA_v^\wedge$ und anschließendes Auffüllen der leeren Einträge gemäß je einer Lösung $TA_w^\vee$ eines oder mehrerer Knoten w mit $(v,w) \in E_{\wedge\to\vee}^{down}$ berechnet werden (siehe Abbildung 7.11).
Da dabei alle möglichen Kombinationen berücksichtigt werden, wird die Anzahl $|TA_v^\wedge|$ der so an einem Knoten $v \in V_\wedge$ erzeugten Lösungen $TA_v^\wedge$ durch die Anzahl der zugehörigen Unterprobleme $SP_w^\vee$ mit $(v,w) \in E_{\wedge\to\vee}^{down}$ (in der Regel zwei) und die Anzahl der verschiedenen Lösungen $TA_w^\vee$ für diese Unterprobleme bestimmt:

$$|TA_v^\wedge| \leq 1 + \prod_{(v,w)\in E_{\wedge\to\vee}^{down}} |TA_w^\vee|$$

Die 1 resultiert aus der Übernahme der partiellen Lösung $PA_v^\wedge$ als Lösung für das Teilproblem, das Produkt steht für die möglichen Kombinationen von $PA_v^\wedge$ mit Lösungen für Unterprobleme, wobei auch die leere Lösung für ein Unterproblem eine erlaubte Lösung darstellt.

7.6.2 Auswahl von Lösungen an Knoten aus $V_\vee$

Die Menge der Lösungen $TA_v^\vee$ an einem Knoten $v \in V_\vee$ ergibt sich durch Vereinigung der Lösungen $TA_w^\wedge$ mit $(v,w) \in E_{\vee\to\wedge}^{down}$ und ist damit so groß wie die Summe der $TA_w^\wedge$. Ohne eine geeignete Einschränkung der Lösungen würde die Anzahl der bei der Zusammenfassung durch Kombination von Unterlösungen erzeugten Lösungen exponentiell mit der Tiefe des Baumes wachsen.
Die Tiefe des Lösungsbaums $\mathcal{T}$ hängt vom bearbeiteten Problem ab. Je entfernter verwandt die betrachtete Sequenz und Struktur sind, desto kürzer sind in der Regel die Abschnitte, die in einen Schritt des rekursiven Abstiegs zugeordnet werden können, desto größer ist die Tiefe des Baumes. Da sowohl für die Anzahl der in einer partiellen Lösung zugeordneten Positionen eine Untergrenze (mindestens 3) gilt, als auch für weiter bearbeitete Teilprobleme eine Mindestgröße (ebenfalls mindestens 3) verwendet wird und zudem eine Aufspaltung selten am Rande eines Teilproblems erfolgt, ist die Tiefe des Baumes in erster Näherung logarithmisch in der Länge der Sequenz beziehungsweise Struktur. Bei den typischerweiser verwendeten Mindestlängen liegt die Tiefe des Baumes $\mathcal{T}$ bei der Lösung von Sequenzstrukturalignmentproblemen für entfernt verwandte beziehungsweise in Faltungserkennungsexperimenten nicht verwandte Sequenzstrukturpaare zwischen 10 und 20, wobei hier nur die Knoten aus $V_\vee$ aus $\mathcal{T}$ gezählt sind. Zieht man außerdem in Betracht, daß die Orakel verschiedene partielle Lösungen (in typischen Anwendungen etwa 12) generieren, wird sofort deutlich, daß es unpraktikabel ist, alle für ein Teilproblem im zugehörigen Teilbaum kodierten Lösungen an die nächsthöhere Hierarchiestufe weiterzureichen. Die daher in der *bottom–up*–Phase notwendige Auswahl von Lösungen erfolgt an Knoten $v \in V_\vee$ durch die Funktion $e_\vee$.
Die Funktion $e_\vee$ besteht in der gegenwärtigen Implementierung der RDP–Methode aus den folgenden nacheinander angewendeten Filtern:

1. Lösche identische Lösungen aus $TA_v^\vee$, die eventuell durch Kombination entstanden sind (siehe Abschnitt 7.3.2).

2. Modifiziere Lösungen aus $TA_v^\vee$ mit Deletionen derart, daß sie zulässig sind (siehe dazu Abschnitt 7.6.2.1).

3. Modifiziere Lösungen aus $TA_v^\vee$ mit Insertionen derart, daß sie zulässig sind (siehe dazu Abschnitt 7.6.2.2).

4. Lösche Lösungen aus $TA_v^\vee$, deren zugehöriges Strukturmodell den durch

Bindungsstelleninformationen oder den Anwender definierten Randbedingungen nicht genügen kann (siehe Abschnitt 7.6.2.3).

5. Bewerte alle verbleibenden Lösungen $TA_v^\vee$ neu (siehe Abschnitt 7.6.2.5) und behalte nur die besten Lösungen (in der Regel 25 oder die aufgrund eines Bonus entsprechend erhöhte Anzahl).

Falls eine Lösung modifiziert wurde, ist sie natürlich erneut auf ihre Zulässigkeit (zum Beispiel bezüglich der Mindestlängen) zu testen, mit den anderen noch gültigen Lösungen aus $TA_v^\vee$ auf Identität hin zu vergleichen und gegebenenfalls zu löschen.

7.6.2.1 Randbedingungen für Deletionen in Lösungen in $\boldsymbol{TA_v^\vee}$

Beim Sequenzstrukturalignment sind Deletionen gleichbedeutend mit Lücken im Rückgrat des aus dem Alignment abgeleiteten Strukturmodells. Russel *et al.* verwenden daher zum Beispiel ein grobes, von der Anzahl der Reste zwischen den Enden zweier Sekundärstrukturelemente abhängiges Abstandskriterium zur Selektion verträglicher Sequenzstrukturpaare aus einer Menge von Paarungen, die durch den Vergleich der für die Sequenz vorhergesagten Sekundärstruktur mit der Sekundärstruktur der Struktur bestimmt werden [299].
Beim Sequenzstrukturalignment mit distanzabhängigen Kontaktpotentialen haben Deletionen jedoch auch eine weitergehende Bedeutung für die Bewertung eines Strukturmodells. So sind die resultierenden Lücken nur zu schließen, indem die Lage der angrenzenden Aminosäurereste verändert wird, denen durch das Alignment noch Reste aus der Sequenz zugeordnet werden. Dies bedeutet jedoch, daß sich die Kontakte der betroffenen Positionen mit hoher Wahrscheinlichkeit ändern. Die Änderungen bezüglich der Kontakte sind natürlich umso gravierender, je feiner die Distanzunterteilung des verwendeten Potentials gewählt wurde.
Daher werden in der RDP–Methode Abstandsbedingungen nicht allein als Ausschlußkriterium für Lösungen verwendet, sondern Lösungen $f \in TA_v^\vee$ werden aufgrund von Abstandsbedingungen modifiziert. Diese Modifikation erfolgt, indem so viele Reste der Sequenz aus dem Alignment entfernt und Insertionen zugeordnet werden, wie mindestens zum Schließen der Schleife zwischen den angrenzenden noch alinierten Positionen benötigt werden.
Die Anzahl der zum Schließen der Schleife benötigten Reste errechnet sich aus der maximal möglichen Entfernung zweier C_α–Atome. Für Peptidbindungen in *trans*–Konformation sind dies 3.75Å und für Peptidbindungen in *cis*–Konformation 2.9Å [285]. Schleifen sind immer kürzer als das Produkt aus dem maximalen C_α–C_α–Abstand (3.75Å) und der Anzahl der Aminosäurereste. Die Anzahl der Reste, die mindestens frei bleiben müssen und nicht Strukturpositionen zugeordnet werden dürfen, errechnet sich daher einfach aus dem Abstand der an die Deletion angrenzenden alinierten Positionen.

Seien i und j zwei Strukturpositionen aus dem aktuellen Teilproblem $SP_v^\vee$ mit $i < j$, $f^{-1}(i) \neq \emptyset$, $f^{-1}(j) \neq \emptyset$ und $\forall_{i<k<j}\ f^{-1}(k) = \emptyset$. Solange der C_α–Abstand der an die Deletion angrenzenden alinierten Positionen i und j größer als 3.75Å mal der um eins erhöhten Anzahl der zwischen $f^{-1}(i)$ und $f^{-1}(j)$ liegenden Reste der Sequenz ist und i und j noch im aktuellen Teilproblem $SP_v^\vee$ liegen, wird entweder i oder j aus der Teillösung entfernt und dekrementiert beziehungsweise inkrementiert. In erster Linie wird dabei der Rest disaliniert, dessen Disalinierung die Differenz zwischen dem C_α–Abstand der an die Deletion angrenzenden alinierten Positionen und der mit den unalinierten Sequenzresten realisierbaren Schleifenlänge minimiert. Falls jedoch der Unterschied kleiner als 1Å ist, wird vorrangig der Reste disaliniert, welcher nicht konserviert und/oder nicht in einen Sekundärstrukturelement liegt.
Diese Modifikationsroutine terminiert in der Regel sehr schnell, da zum Beispiel mit drei Resten (vier C_α–Abstände) nach diesem Kriterium bereits ein Abstand von 15Å überbrückt werden kann. In jedem Fall terminiert sie, wenn die Teillösung zur leeren Lösung geworden ist, das heißt keine Zuordnung mehr stattfindet.
Obwohl das verwendete Kriterium nicht sehr hart ist, wird durch die beschriebene Modifikation von Lösung effizient verhindert, daß Aminosäurereste in die Kontaktpotentialbewertung einbezogen werden, die definitiv in dem Strukturmodell die zur Bewertung der ursprünglichen Lösung verwendeten Kontakte nicht ausbilden können.

7.6.2.2 Randbedingungen für Insertionen in Lösungen in $TA_v^\vee$

Ähnlich wie bei Deletionen gibt es auch für Insertionen bezüglich der verwendeten Struktur gewisse Randbedingungen, die erfüllt sein sollten, damit diese in dem angestrebten Strukturmodell realisiert werden können. So sind Insertionen unter Beibehaltung der Lage der restlichen Strukturelemente — und damit der zugehörigen Kontakte — nur dann möglich, wenn sie an Stellen in der Struktur stattfinden, die ein Ausweichen der Proteinkette in den freien Raum (beziehungsweise das Lösungsmittel) erlauben.
Daher werden in der RDP–Methode Lösungen $f \in TA_v^\vee$ mit Insertionen ähnlich wie bei Deletionen auf ihre Realisierbarkeit hin überprüft und, wenn möglich, so modifiziert, daß die Proteinkette in dem resultierenden Strukturmodell in das Lösungsmittel ausweichen kann und die Kontakte der alinierten Strukturpositionen erhalten bleiben.
Als Kriterium wird dabei die Lösungsmittelzugänglichkeit der Strukturpositionen verwendet, zwischen denen die Insertion stattfindet. Die dem Lösungsmittel zugängliche Fläche einer Strukturposition wird bei der Verwendung von Distanzkontaktrelationen der `DSSP`–Datenbank [173] entnommen. Beim Einsatz von Voronoikontaktrelationen wird die Kontaktfläche zu Gitterpunkten einer Position (siehe dazu Abschnitt 5.2.2) als Maß für die Zugänglichkeit verwendet.

Für eine zulässige Insertion wird gefordert, daß die zwei angrenzenden Strukturpositionen beide eine dem Lösungsmittel zugängliche Fläche von mindestens 25Å^2 (in Anlehnung an die halbe Kugeloberfläche eines Kohlenstoffatoms) haben, um so Insertionen auch an Positionen mit sehr kleiner Lösungsmittelzugänglichkeit auszuschließen.
Wenn beide Strukturpositionen hinreichend lösungsmittelzugänglich sind, wird die Lösung f als zulässig anerkannt und zusätzlich die darin für die Positionen getroffene Zuordnung aufgehoben, da durch die Insertion voraussichtlich auch die räumliche Lage der beiden Reste im resultierenden Strukturmodell beeinflußt wird. Lösungen mit nicht zulässigen Insertionen werden direkt verworfen.
Ein Sonderfall tritt dann ein, wenn es sich nicht um eine einfache Insertion, sondern um ein sogenanntes Wechselgap handelt. Ein Wechselgap liegt dann vor, wenn gleichzeitig eine Insertion und eine Deletion durch die Lösung vorgegeben werden. Falls die Insertion kürzer als die Deletion ist, wird das Wechselgap wie eine Deletion behandelt (siehe Abschnitt 7.6.2.1). Falls die Insertion länger ist, steht nicht so sehr die Bedingung der Schließbarkeit der Proteinkette im Vordergrund, vielmehr ist zu beachten, daß die zusätzlichen Aminosäurereste in das Strukturmodell eingebracht werden können. Für die Zulässigkeit einer Lösung f ist es in diesem Falle hinreichend, wenn mindestens zwei der Strukturpositionen, denen durch f keine Aminosäure der Sequenz zugeordnet ist, eine Lösungsmittelzugänglichkeit aufweisen, die größer als die angegebene Mindestzugänglichkeit für eine Insertion ist. Falls dies der Fall ist wird die Lösung f als zulässig angesehen und nicht modifiziert, ansonsten wird verfahren, wie dies oben für normale Insertionen beschrieben wurde.
Die für Deletionen und Insertionen verwendeten Randbedingungen sind natürlich nur notwendige aber keine hinreichenden Bedingungen für die Realisierbarkeit von aus Insertionen und Deletionen resultierenden Schleifen. Zum Beispiel wird nicht berücksichtigt, ob nicht andere Reste des Strukturmodells die Realisierung der Schleife blockieren oder verhindern. Daher soll im Anschluß an diese Arbeit ein auf Wegemethoden basierendes Verfahren entwickelt werden (siehe auch Abschnitt 9.3), das erstens die Realisierbarkeit der neu zu modellierenden Schleifen in dem aus der jeweiligen Lösung resultierenden Strukturmodell überprüft und zweitens einen optimalen Weg durch das von den bereits zugeordneten Positionen induzierte Kontaktpotentialfeld sucht.

7.6.2.3 Bindungsstelleninformation zur Einschränkung von $\boldsymbol{TA}_v^{\vee}$

In der RDP–Methode wird Bindungsstelleninformation nicht nur, wie bereits beschrieben, als mögliches Orakel verwendet, sondern aus Bindungsstelleninformationen werden auch Randbedingungen abgeleitet.
Im einfachsten Fall kann für die Aminosäurereste der Bindungsstelle gefordert werden, daß sie in dem aus der Teillösung resultierendem Strukturmodell dem Lösungsmittel und damit dem potentiellen Liganden zugänglich sein müssen. Eine

Teillösung, in der diese Reste Strukturpositionen zugeordnet werden, ist damit nur dann zulässig, wenn die Positionen dem Lösungsmittel zugänglich sind.
Ist nur die Funktion des Proteins bekannt, kann daraus in einigen Fällen abgeleitet werden, daß bestimmte Aminosäuretypen in der Struktur in einer bestimmten räumlichen Anordnung auftreten müssen, aber es muß nicht unbedingt bekannt sein, welche Reste dies in der Sequenz sind. Zum Beispiel kann man aus einer Serinproteaseaktivität zwar in der Regel auf die Existenz einer katalytischen Triade (siehe Abbildung 4.8) schließen, aber nicht immer sind alle an der Ausbildung der Triade beteiligten Reste aus der Sequenz eindeutig bestimmt. Für ein Alignment, beziehungsweise das daraus resultierende Modell, kann überprüft werden, ob es im Modell entsprechende Aminosäuretypen in einer entsprechenden Anordnung gibt. Leider ist dieser Test erst dann möglich, wenn alle zur Auswahl stehenden Aminosäurereste der Sequenz Bestandteil einer Teillösung sind. Die Entscheidung über die Zulässigkeit ist damit im schlechtesten Fall erst mit dem Erreichen der Wurzel des Lösungsbaums in der *bottom–up*–Phase möglich.
Denkbar sind auch Randbedingungen [246], die fordern, daß bestimmte Aminosäurereste der Sequenz zu einer Position *vor* oder *nach* einer bestimmten Strukturposition zuzuordnen sind. Dies ist zum Beispiel dann der Fall, wenn aus Experimenten bekannt ist, daß sich ein bestimmter Rest in einer bestimmten Domäne eines aus mehreren Domänen bestehenden Proteins befindet. Da die entsprechende Information in der Annotationsliste der Sequenzposition abgelegt ist, kann die Zulässigkeit einer Teillösung, die den entsprechenden Aminosäurerest enthält, einfach überprüft werden.

7.6.2.4 Weitere Randbedingungen zur Einschränkung von $TA_v^\vee$

Neben Bindungsstelleninformation können auch andere diskrete Zusatzinformationen wie zum Beispiel experimentell bestimmte Schwefelbrücken oder geladene/ polare Aminosäuren ohne geeignete Wechselwirkungspartner im hydrophoben Kern eines Strukturmodells zur Filterungen von Lösungen aus $TA_v^\vee$ verwendet werden.
Falls bekannt ist, welche Schwefelbrücken zwischen Struktur und untersuchter Sequenz konserviert sind, wird die Zuordnung der entsprechenden Cysteine als Orakel verwendet. Diese Art Orakel erzeugt bereits an der Wurzel des Lösungsbaums $\mathcal{T}$ eine sicherer Zerlegung des Gesamtproblems. Falls es sich um nicht konservierte Schwefelbrücken handelt, kann ihre Existenz in einem Teilstrukturmodell und damit die zugehörige Lösung aus $TA_v^\vee$ nur getestet werden, wenn beide an der Schwefelbrücke beteiligten Cysteine bereits im Teilstrukturmodell vorhanden sind.
Das Kriterium, daß in einem sinnvollen Strukturmodell keine geladenen oder polaren Aminosäuren ohne geeignete Wechselwirkungspartner im hydrophoben Kern auftreten sollten, kann ebenfalls erst angewendet werden, wenn das Teilstrukturmodell auch die zu der in Frage stehenden Position benachbarten Positionen

umfaßt.
Die Flexibilität der RDP–Methode erlaubt zudem die Einbeziehung weiterer durch den Benutzer definierter Randbedingungen. Die Entwicklung einer Regelsprache und eines Regeleditors, die es einem biologisch oder biochemisch orientierten Benutzer erlauben, über die Sequenz bekanntes Wissen in Regeln zur Steuerung der Berechnung des Sequenzstrukturalignments umzusetzen, ist Gegenstand zukünftiger Forschung (siehe Abschnitt 9.3).

7.6.2.5 Neubewertung der Lösungen aus $TA_v^{\vee}$

Eine Neubewertung der Teillösungen in der *bottom–up*–Phase ist aus mehreren Gründen notwendig: Zum einen sind durch die Kombination von Teillösungen und ihre Modifikation zur Erfüllung von Randbedingungen neue bisher nicht bewertete Teillösungen entstanden, zum anderen ist in der *bottom–up*–Phase weit mehr über die Gesamtlösung und damit über die Kontaktpartner der Reste des Teilproblems im resultierenden Strukturmodell bekannt, als dies in der *top–down*–Phase der Fall ist.
Nachdem alle nicht zulässigen Lösungen aus der Kandidatenmenge gelöscht wurden, erfolgt die Neubewertung der Teillösungen anhand der folgenden Kostenfunktion:

$$\begin{aligned}\phi(f', A', B') \;=\; & \gamma * \phi^S(f', A', B') \;+\; \delta * \phi^C(f', A', B') \;+ \\ & \epsilon * \phi^H(f', A', B') \;+\; \zeta * \phi^1_{pot}(f', A', B') \;+ \\ & GAP(f', A', B')\end{aligned}$$

Diese Kostenfunktion ist identisch mit der, die bei Orakeln mit gemischter Kostenfunktion verwendet wird (siehe Abschnitt 7.2.3.5). Die Kalibrierung der Parameter γ, δ, ϵ und ζ wird im folgenden Abschnitt 7.7 beschrieben.

7.7 Parameterkalibrierung

Bei Orakeln mit verschiedenen Kostenfunktionsanteilen (siehe Abschnitt 7.2.3.5) bei der Neubewertung von Teillösungen (siehe Abschnitt 7.6.2.5) und bei der abschließenden Bewertung eines berechneten Sequenzstrukturalignments gehen verschiedene Terme in die Kostenfunktion ein. Die Gewichtung der Kostenfunktionsanteile untereinander ist ein derzeit nicht allgemein lösbares Problem. Daher erfolgt die Kalibrierung der Parameter γ, δ, ϵ und ζ empirisch anhand einer ausgewählten Datenmenge.
Als erste Näherung werden die Gewichtungen so eingestellt, daß die Kostenfunktionsanteile ungefähr in der gleichen Größenordnung liegen. Da die RDP–Methode mit verschiedenen Potentialen und anderen Bewertungskriterien arbeitet, muß eigentlich bei jedem Wechsel in einer der Komponenten der Bewertungsfunktion eine Rekalibrierung erfolgen. Eine empirische Kalibrierung hat jedoch

auf einer umfangreichen Datenbasis zu erfolgen und kann daher nicht immer neu durchgeführt werden. Außerdem sind *die* optimalen Parameter sicher auch vom jeweiligen Anwendungsfall und der verfolgten Zielvorgabe abhängig. So sollte sicher dem Sequenzanteil beim strukturgenauen Alignment von evolutionär verwandten Proteinen mehr Bedeutung beigemessen werden, als wenn es um die Erkennung struktureller Verwandtschaften zwischen Proteinen geht, deren Verwandtschaft erst auf struktureller Ebene erkennbar ist.
Die Suche nach *der* optimalen Einstellung der Parameter der RDP–Methode bleibt, wie auch die Auswahl und Ableitung noch besserer Potentiale, die Aufgabe zukünftiger Untersuchungen. Die im folgenden diskutierten Ergebnisse geben nur erste Hinweise auf die optimale Parameterwahl.

7.7.1 Empirische Methode zur Einstellung der Parameter

Die RDP–Methode zielt hauptsächlich auf die Berechnung strukturgenauer Sequenzstrukturalignments für Sequenzstrukturpaarungen, wo die herkömmlichen sequenzbasierten Verfahren scheitern. Die Problematik der Bewertungskriterien für die Güte von Sequenzstrukturalignments wird in Abschnitt 8.3.1 diskutiert werden. Die folgenden Untersuchungen finden an Proteinen statt, deren Strukturen aufgeklärt sind. Daher wird die *RMS*–Abweichung der strukturellen Superposition gemäß des Alignments als Kriterium für die Güte eines Sequenzstrukturalignments verwendet.
Das nächste Problem bei der empirischen Parameterkalibrierung liegt in der Wahl der Beispielmenge, auf der die Kalibrierung durchgeführt wird. Da es sehr schwierig ist, geeignete Testmengen (siehe Abschnitt 8.3.2) zu finden, wird hier im folgenden auf die Menge der mit dem Programm `SARF2` berechneten Strukturalignments zurückgegriffen, um den Einfluß der unterschiedlichen Gewichtungsfaktoren nachzuweisen. Eine detaillierte Beschreibung der Eigenschaften dieser Strukturalignments wird in Abschnitt 8.3.2 gegeben.
Der Effekt der verschiedenen Parameter wird an der Zahl der Beispiele gemessen, für die die RDP–Methode mit den gegebenen Parametern Sequenzstrukturalignments mit *RMS*–Abweichungen berechnet, die maximal 2Å über der des besten für das jeweilige Strukturpaar mit `SARF2` berechneten Strukturalignments liegen. Daß dies für Sequenzstrukturalignmentmethoden ein sehr hartes Gütekriterium ist, zeigt der Vergleich mit anderen Methoden in Abschnitt 8.3.
Durch die empirische Parameterkalibrierung sollen nun die Gewichtungsparameter γ, δ, ϵ und ζ so eingestellt werden, daß die Anzahl der Sequenzstrukturalignment mit niedriger *RMS*–Abweichung maximiert wird. Der Sequenzanteil γ wird dazu konstant auf 1.0 gesetzt und die anderen Gewichtungsparameter werden (mehr oder weniger) systematisch variiert. Da eine systematische Analyse aller Gewichtungsparameter den Rahmen dieser Arbeit sprengen würde, wird schwerpunktmäßig die Gewichtung ζ des Paarpotentialanteils bei Verwendung unterschiedlicher empirischer Paarpotentiale untersucht.

Für jeden Meßpunkt in den im folgenden Abschnitt diskutierten Abbildungen 7.12 bis 7.14 werden also bei jeweils fester Gewichtung, fester Einstellung sonstiger Parameter und festem Paarpotential mit der RDP–Methode 73 Sequenzstrukturalignments und ihre *RMS*–Abweichung berechnet. Wie gut ein Parametersatz ist, wird anhand der Anzahl der Sequenzstrukturalignments entschieden, deren *RMS*–Abweichung weniger als 2Å über der des Strukturalignments liegt. Die Gewichtung des Paarpotentialanteils der Kostenfunktion wird dabei mit einer Schrittweite von 2.0 im Intervall von 0.0 bis 30.0 variiert. Die anderen Gewichtungen und Parameter sind für jede der Kurven fest und wie jeweils angegeben gewählt. Für jede der gezeigten Kurven sind also mit der RDP–Methode 1168 Sequenzstrukturalignments berechnet worden.

7.7.2 Vergleich verschiedener Parametersätze

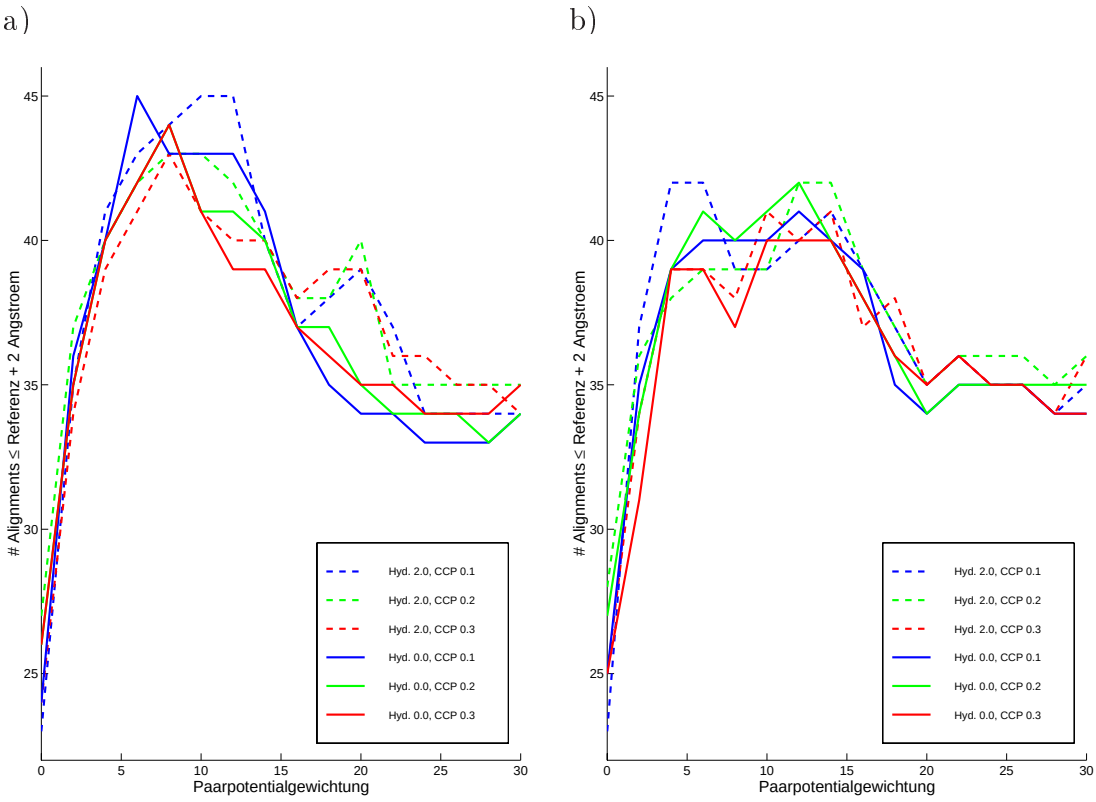


Abbildung 7.12: Güte von 73 *free–shift*–Sequenzstrukturalignments bei Gewichtungen $\zeta = 0, \ldots, 30$ für die Paarpotentiale a) `ED6SD6` und b) `ED3SD2` (beide auf Voronoikontaktrelationen).

Abbildung 7.12 zeigt sehr deutlich den Einfluß, den die Verwendung eines Paarpotentialterms in der Kostenfunktion auf die Güte der Alignments hat. Das für Teil a) von Abbildung 7.12 verwendete Potential basiert auf Voronoikontaktrela-

tionen und differenziert Kontakte in sechs Distanz– und sechs Sequenzabstandsintervalle. Die Gewichtung des Paarpotentialanteils geht in 16 Schritten von 0.0 bis 30.0. Ohne Paarpotential ($\zeta = 0.0$) werden in Abhängigkeit von den anderen Parametern nur für 23 bis 26 Paare Sequenzstrukturalignments mit einer *RMS*–Abweichung kleiner als 2Å über der *RMS*–Abweichung des Strukturalignments erzielt. Bei einer Paarpotentialgewichtung ζ zwischen 6.0 und etwa 12.0 erreichen dagegen mindestens 40 der Alignments dieses Qualitätskriterium.
Für die besten Parameterkombinationen in diesem Bereich liegen sogar 45 Alignments unterhalb der gesetzten Grenze. Die besten Alignments werden bei einer CCP–Gewichtung von $\delta = 0.1$ entweder mit einer Hydrophobizitätsgewichtung $\epsilon = 0.0$ und einer Paarpotentialgewichtung von $\zeta = 8.0$ oder mit einer Hydrophobizitätsgewichtung $\epsilon = 2.0$ für ζ zwischen 10.0 und 12.0 erzielt.
Mit größerem Paarpotentialgewicht nimmt die Alignmentqualität ab, sinkt aber nicht mehr unter die Marke von 33 guten Alignments ab. Stichprobenartige Tests mit Gewichten $\zeta > 30.0$ haben gezeigt, daß jenseits dieser Grenze keine besseren Ergebnisse mehr zu erwarten sind, aber erstaunlicherweise die Alignmentqualität nicht mehr gravierend abnimmt, solange die Gewichte im normalen Rahmen bleiben. Vermutlich wird dieses Verhalten zum einen durch die zusätzlichen Randbedingungen für Insertionen und Deletionen, die bei der RDP–Methode einfließen, und zum anderen dadurch bewirkt, daß nur der vernünftige Teil des für Sequenzstrukturalignments möglichen Suchraumes durch die partiellen Lösungen aus den verschiedenen Orakeln abgedeckt wird.
Wie ein Vergleich der gestrichelten ($\epsilon = 2.0$) und der durchgezogenen ($\epsilon = 0.0$) Linien zeigt, hat das Hydrophobizitätspotential keinen eindeutigen Einfluß auf die Alignmentqualität. Die Ergebnisse mit Hydrophobizitätsanteil scheinen geringfügig besser zu sein. Da aber das angewendete Qualitätskriterium eine harte Grenze bei 2.0Å plus RMS–Abweichung des Strukturalignments setzt und die Anzahl der alinierten Positionen nicht betrachtet wird, kann dies auch ein Artefakt des Gütekriteriums sein.
Die verschiedenen Farben kodieren verschiedene Gewichtungen $\delta = 0.1, 0.2$ und 0.3 des CCP–Potentialterms. Die Spitzenwerte werden mit niedrigem δ erreicht. Aber auch hier läßt sich kein eindeutiger Trend erkennen. Der wesentliche Gewinn durch die Verwendung von CCP–Potentialen scheint in der Generierung sinnvoller partieller Lösungen in den Orakelschritten der RDP–Methode zu liegen, wo die anderen Ähnlichkeitsmaße versagen. Darauf deutet zumindest die Tatsache hin, daß die Ergebnisse schlechter werden, wenn man kein CCP–Potential verwendet.
Abbildung 7.12 b) zeigt das gleiche Experiment für ein auf Voronoikontaktrelationen basierendes Paarpotential, das Kontakte nur in drei Distanz– und zwei Sequenzabstandsintervalle unterscheidet. Ohne Paarpotential ($\zeta = 0.0$) werden hier in Abhängigkeit von den anderen Parametern 23 bis 28 Sequenzstrukturalignments mit einer *RMS*–Abweichung kleiner als 2Å über der *RMS*–Abweichung des Strukturalignments berechnet. Bei einer Paarpotentialgewichtung ζ zwischen 6.0 und etwa 12.0 erreichen dagegen mindestens 37 der Alignments dieses Qualitäts-

kriterium. Mit größerem Paarpotentialgewicht nimmt die Alignmentqualität zwar ab, sinkt jedoch nicht mehr unter 34 gute Alignments.
Ingesamt sind damit die Ergebnisse, was die Alignmentqualität anbelangt, für das feinere Potential `ED6SD6` (Abbildung 7.12 a)) eindeutig besser als für das grobere Paarpotential `ED3SD2` (Abbildung 7.12 b)). Dagegen scheint die optimale Gewichtung der Parameter durchaus vergleichbar zu sein.

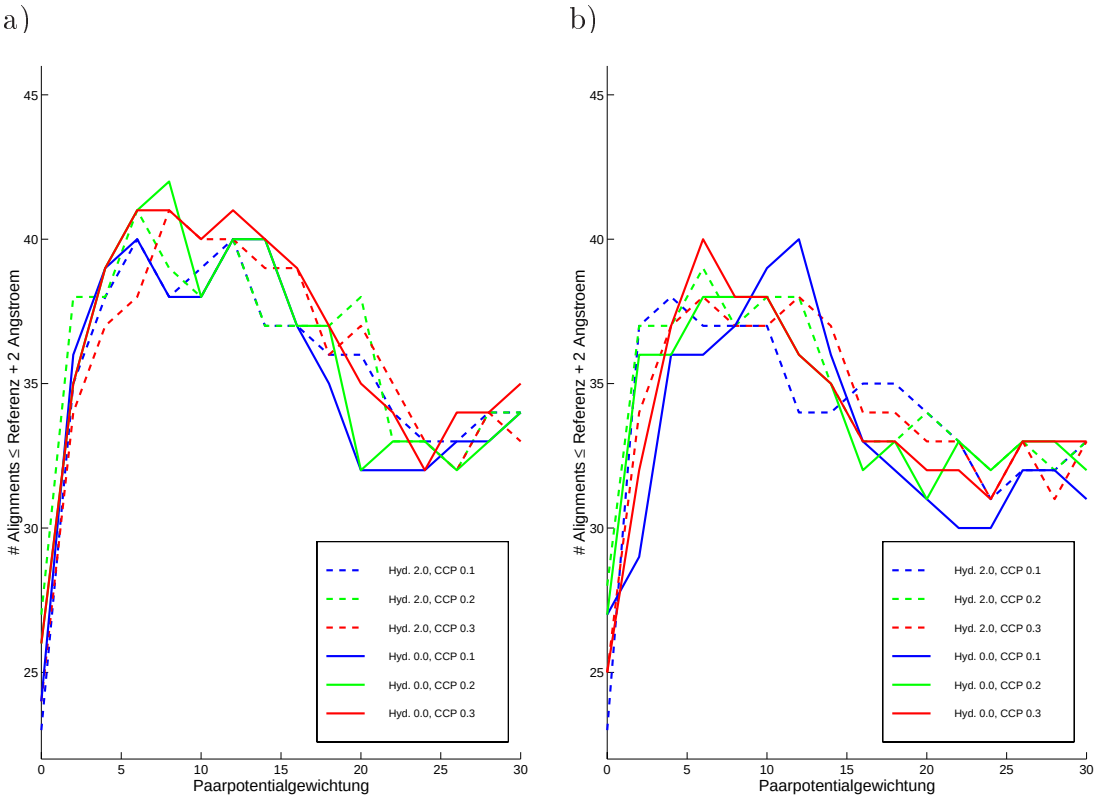


Abbildung 7.13: Güte von 73 globalen Sequenzstrukturalignments bei Gewichtungen $\zeta = 0, \ldots, 30$ für die Paarpotentiale a) `ED6SD6` und b) `ED3SD2` (beide auf Voronoikontaktrelationen).

Abbildung 7.13 zeigt die Daten für das gleiche Experiment wie Abbildung 7.12 mit dem Unterschied, daß hier globale Sequenzstrukturalignments berechnet wurden. Wie nicht anders zu erwarten, liegt die *RMS*–Abweichung der globalen Sequenzstrukturalignments etwas höher als die der *free–shift*–Alignments, da bei einem globalen Alignment tendentiell mehr Reste und damit in der Regel auch die zwischen den Strukturen unterschiedlichen Anfangs– und Endbereiche aliniert werden, die in *free–shift*–Alignments unaliniert bleiben.
Aber auch bei der Berechnung globaler Alignments ist die Alignmentqualität bei Verwendung des feineren Potentials `ED6SD6` (siehe Abbildung 7.13 a)) besser als bei Verwendung des groberen Potentials `ED3SD2` (siehe Abbildung 7.13 b)). Für das bessere Potential erreichen mit der besten Parameterkombination

($\gamma = 1.0, \delta = 0.2, \epsilon = 0.0$ und $\zeta = 8.0$) immerhin 42 und für viele andere Kombinationen mehr als 40 der Sequenzstrukturalignments das gewählte Kriterium. Dagegen liefern für das Potential ED3SD2 nur die besten Parameterkombinationen 40 gute Sequenzstrukturalignments.
Wie auch bei den *free–shift*–Alignments liegt die optimale Gewichtung ζ für den Paarpotentialanteil in der Kostenfunktion im Bereich zwischen 6.0 und 16.0 für ED6SD6 und zwischen 6.0 und 12.0 für ED3SD2. Dies belegt, daß die optimale Gewichtung des Paarpotentialanteils hauptsächlich vom Potential und weniger vom Alignmentmodus abhängt.
Die zusätzliche Verwendung eines Hydrophobizitätspotentials bedeutet auch hier nur in wenigen Fällen einen Zugewinn an Alignmentqualität und die besten Ergebnisse werden sogar mit $\epsilon = 0.0$ erzielt. Der Grund dafür liegt wahrscheinlich darin, daß sowohl das CCP–Potential als auch das Paarpotential indirekt auch Anteile der Hydrophobizität kodieren. Daher werden die folgenden Experimente ohne einen zusätzlichen Hydrophobizitätsterm durchgeführt.
Über die dargestellten Ergebnisse hinausgehende Untersuchungen haben gezeigt, daß sich alle auf Voronoikontaktrelationen basierenden Potentiale [366], die Kontakte nach der euklidischen Distanz und dem Sequenzabstand der Kontaktpartner unterscheiden, ähnlich wie die hier näher betrachteten Potentiale verhalten. Daher wird auf eine Präsentation der Ergebnisse für Potentiale ähnlichen Typs verzichtet. Dies gilt aber nur solange die Unterscheidungen der Kontakte nicht zu fein werden. Bereits die Ergebnisse für ein Potential ED12SD6, in dem Kontakte in zwölf Distanzintervalle unterschieden werden, sind schlechter als für das hier betrachtete ED6SD6. Der Grund dafür liegt – wie es auch die Ergebnisse von Halfmann [131] andeuten – darin, daß sich die Kontakte in entfernt verwandten Proteinen, deren Verwandtschaft nur in einer ähnlichen Struktur besteht, hinsichtlich der Distanz der kontaktierenden Reste recht stark unterscheiden.
Die von Sippl [319] geäußerte Vermutung, daß eine Ursache für das schlechtere Abschneiden detaillierter Potentiale die geringere Anzahl der Zählungen sein könnten, trifft hier nicht zu, da zur Vermeidung statistischer Probleme bei allen hier betrachteten Potentialen die in der eigentlichen Struktur beobachteten Kontakte zwischen bestimmten Aminosäuretypen durch Hinzunahme der Reste von zu der Struktur homologen Proteinen aus der HSSP–Datenbank angereichert werden und so in allen Intervallen genügend Ereignisse gezählt werden.
Auf eine detaillierte Diskussion anderer Potentiale ähnlichen Typs wird daher zugunsten der Analyse von Potentialen verzichtet, die auf Distanzkontaktrelationen basieren oder die Kontaktfläche mit einbeziehen.
Abbildung 7.14 a) zeigt zum Vergleich die Qualität von Sequenzstrukturalignments, die mit RDP unter Verwendung eines herkömmlichen, auf Distanzrelationen basierenden Potentials berechnet wurden. Aus Gründen der Vergleichbarkeit wurde dabei das Potential ED3SD2 gewählt, dessen einziger Unterschied zu dem in Abbildungen 7.12 b) und 7.13 b) in der Kontaktdefinition besteht. Unabhängig davon, ob globale (gestrichelt) oder *free–shift*–Alignments (durchge-

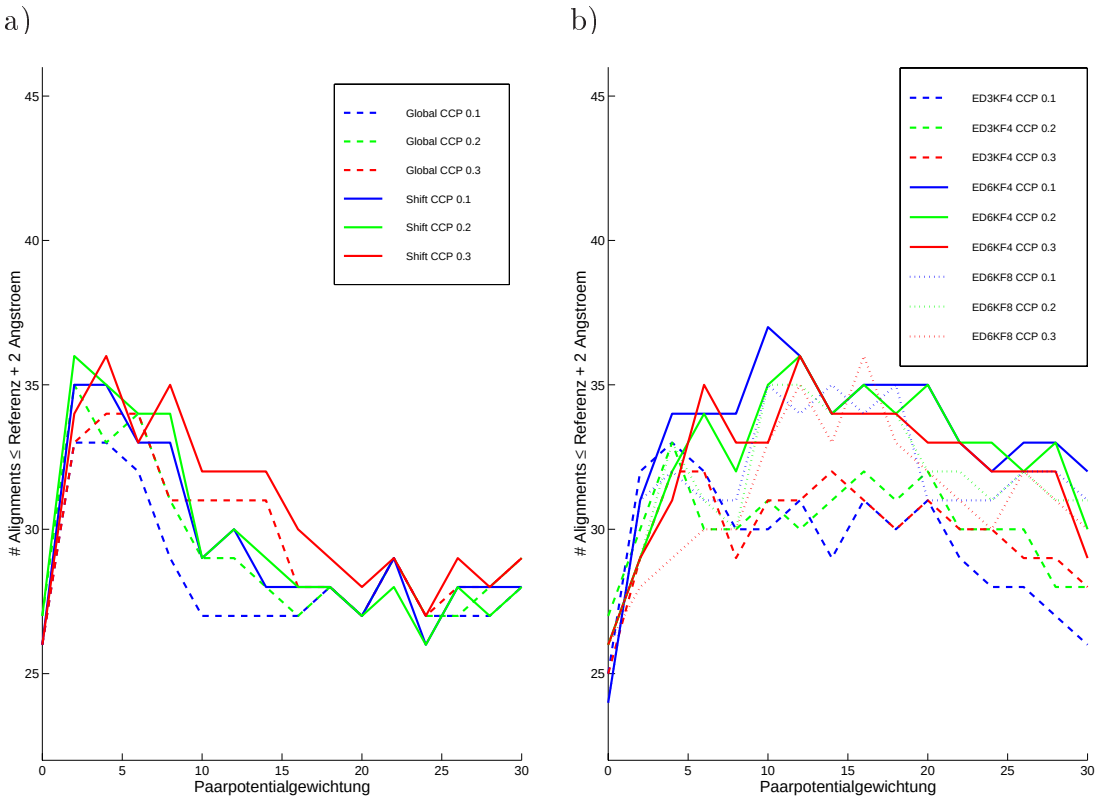


Abbildung 7.14: Güte a) von 73 globalen (gestrichelt) und 73 *free–shift*–Sequenzstrukturalignments (durchgezogen) für ein Paarpotential `ED3SD2` auf Distanzkontaktrelationen und b) von 73 *free–shift*–Sequenzstrukturalignments für kontaktflächenabhängige Paarpotentiale `ED3KF4` (gestrichelt) `ED6KF4` (durchgezogen) und `ED6KF8` (gepunktet) auf Voronoikontaktrelationen für Potentialgewichtung $\zeta = 0, \ldots, 30$.

zogen) berechnet werden, ist das Distanzkontaktpotential in zweifacher Hinsicht schlechter als das auf Voronoikontakten basierende Potential. Erstens erreichen selbst für die beste Parameterwahl nur 36 Alignments das Qualitätskriterium, während dies bei der Verwendung des vergleichbaren Voronoipotentials immerhin für 42 Alignments erreichbar ist. Und zweitens ist auch das Fenster der Paarpotentialgewichtung ζ, für das der Paarpotentialterm einen positiven Einfluß auf die Alignmentqualität hat, sehr viel enger und liegt etwa zwischen 2 und 6.

Für Abbildung 7.14 b) wurde nach der besten Paarpotentialgewichtung für verschiedene, auf Voronoikontaktrelationen basierende Potentiale gesucht, die Kontakte anstelle nach dem Abstand der Kontaktpartner in der Sequenz nach der Summe der Voronoiflächen unterscheiden, die die kontaktierenden Reste gemeinsam haben. Da die vorangegangenen Tests gezeigt haben, daß sich globale und *free–shift*–Alignments in etwa gleich verhalten, werden hier nur die *free–shift*–

Alignments betrachtet. Stattdessen werden verschiedene kontaktflächenabhängige Potentiale betrachtet, für die sowohl die Einteilung der Abstandsintervalle als auch der Kontaktflächenintervalle variiert wurde. Auch hier werden die besten Ergebnisse (maximal 37 gute Alignments) mit einem Potential mit mittlerer Granularität (`ED6KF4`) erreicht (siehe Abbildung 7.14 b) durchgezogen). Eine Verfeinerung der Einteilung der Flächen (`ED6KF8`) (gepunktet) bringt keinen zusätzlichen Gewinn. Durch Wahl einer groberen Einteilung der Abstandsintervalle werden die Ergebnisse deutlich schlechter (gestrichelt).
Aufgrund der in diesem Abschnitt durchgeführten Experimente wird für die in Kapitel 8 vorgestellten Ergebnisse das auf Voronoikontakten basierende Potential `ED6SD6` verwendet, sofern dies nicht anderweitig spezifiziert wird. Dabei wird der Sequenzterm ϕ^S in der Kostenfunktion ϕ mit $\gamma = 1.0$, der `CCP`–Term ϕ^C mit $\delta = 0.1$, der Hydrophobizitätsterm ϕ^H mit $\epsilon = 2.0$ und der Paarpotentialanteil ϕ^1_{pot} mit $\zeta = 10.0$ gewichtet.
Die `CCP`–Potentiale sind so normiert, daß ihr Anteil an der Kostenfunktion in etwa dem der Sequenz entspricht [379]. Der Faktor 0.1 bedeutet daher eine nicht unwesentliche Abgewichtung des `CCP`–Anteils. Dagegen wird der Anteil des Paarpotentials, das Kontakte mit zwischen -5.0 und $+5.0$ – mit Gewichtung also zwischen -50 und $+50$ – bewertet, gegenüber der Sequenzinformation stark hochgewichtet, da typischerweise die Einträge der Austauschmatrizen zwischen -10 und $+25$ liegen.

Kapitel 8
Ergebnisse

8.1 Definition der Erfolgskriterien

Die vorgestellte RDP–Sequenzstrukturalignmentmethode verfolgt zwei Ziele:

- die Verbesserung der Vorhersage entfernter struktureller Verwandtschaften (*Faltungserkennung*).
- die Verbesserung von Sequenzstrukturabbildungen bezüglich ihrer Verwendbarkeit als Startpunkt für die vergleichende Modellierung (*Alignmentqualität*).

Eine Evaluierung der vorgestellten Methode erfolgt daher auch anhand dieser beiden Zielvorgaben. Man würde erwarten, daß das Erkennen entfernter Sequenzstrukturbeziehungen eine hohe Qualität der Sequenzstrukturabbildung voraussetzt. Daß dies nicht immer der Fall sein muß, hat bereits die Auswertung des ersten Strukturvorhersagewettbewerbes (CASP I) [197] gezeigt. Dies ist umso überraschender, da in der Regel für die Erkennung Bewertungssysteme verwendet wurden, die die Strukturen einer Repräsentativmenge anhand der Abbildung der untersuchten Sequenz in diese Strukturen bewerten und erkennen.
Zur Beurteilung, inwieweit die RDP–Methode die gestellten Zielvorgaben erreicht, gibt es im wesentlichen drei Möglichkeiten:

- im Vergleich mit Verfahren, die das Vorhandensein struktureller Verwandtschaften anhand experimentell aufgeklärter Strukturen nachweisen,
- im Wettbewerb mit anderen Vorhersagemethoden durch Vergleich der Faltungserkennungsrate und Alignmentqualität,
- durch echte Blindvorhersagen mit anschließender Bewertung der Vorhersage im Vergleich der Vorhersage mit der experimentell aufgeklärten Struktur.

Da das Problem der (automatischen) Proteinstrukturvorhersage auch mit der RDP–Methode nicht endgültig gelöst wird, kann naturgemäß auch die RDP–Methode nicht die durch den strukturellen Vergleich gesteckten Ziele in allen Fällen erreichen. Daher ist es auch in diesen Fällen sinnvoll, einen Vergleich mit anderen Methoden durchzuführen.
Es ist klar, daß echte Blindvorhersagen nicht nur der realen Anwendung von Vorhersagemethoden am nächsten kommen, sondern auch den ultimativen Test darstellen. Zudem stellt das Aufdecken neuer Erkenntnisse für denjenigen, der eine Vorhersage erstellt, einen wesentlich größeren Anreiz dar als die Reproduktion

bereits bekannter Resultate. Blindvorhersagen beinhalten in der Regel sowohl die Erkennung der Verwandtschaft der untersuchten Sequenz zu einer oder mehreren bekannten Strukturen als auch die Berechnung eines genauen Alignments der Sequenz mit einer Struktur zum Zwecke der Generierung einer Modellstruktur.
Während der Implementierung der RDP–Methode und der Erstellung dieser Arbeit gab es zwei Gelegenheiten zu echten Blindvorhersagen mit der Möglichkeit zum anschließenden Vergleich gegen die experimentell bestimmten Proteinstrukturen.
Im ersten Fall handelt es sich um eine Einzelvorhersage für die Struktur der Thymidinkinase des *Herpes Simplex Virus I*. Abschnitt 8.6 diskutiert sowohl die Vorgehensweise als auch die erzielte Vorhersage im Vergleich mit der zwischenzeitlich aufgeklärten Struktur.
Im zweiten Fall handelt es sich um die Vorhersagen, die von der Projektgruppe `PROTAL` im Rahmen des internationalen Wettbewerbs für Methoden für die Vorhersage von Proteinstrukturen *Critical Assessment of Protein Structure Prediction Methods* II (CASP II) eingereicht wurden. Abschnitt 8.5 diskutiert die für diesen Wettbewerb eingereichten Vorhersagen. Da insbesondere die RDP–Methode zum Zeitpunkt des Wettbewerbs nicht den heutigen Entwicklungsstand aufwies, wird in Ergänzung zu den damaligen Vorhersagen anhand von einzelnen Beispielen diskutiert, wie eine mögliche Vorhersage heute aussehen würde. Dabei zeigt sich insbesondere, daß die Qualität der Alignments mit Hilfe der RDP–Methode heute weitaus höher ist, als dies mit den zum Zeitpunkt des Wettbewerbs verfügbaren Methoden möglich war. Es ist jedoch zu betonen, daß es sich bei diesen nachwettbewerblichen Verbesserungen nicht mehr um echte Blindvorhersagen handelt. Dies gilt auch dann, wenn wie hier die Kenntnis der Struktur nur zur Analyse und nicht zur Berechnung der Ergebnisse oder zur Kalibrierung der Methode verwendet wird.

8.2 Der ToPLign–Ansatz für Blindvorhersagen

Die Erstellung echter Strukturvorhersagen geht über die Anwendung einer einzelnen Methode hinaus. Bei den in dieser Arbeit beschriebenen Vorhersagen wurden zum Beispiel folgende Einzelschritte angewendet:

- Sequenzsuche in den existierenden Datenbanken mit Standardsuchverfahren wie `BLAST` [10], `FASTA` [207, 274] oder eine Datenbanksuche mit einem lokalen Alignmentverfahren (`BLITZ` [330]).
- Suche nach vorhandener funktioneller Information über die Sequenz:
  - Literaturrecherche (zum Beispiel in `Medline` über `Entrez` [286]),
  - Suche nach funktionellen Sequenzmustern (zum Beispiel `PROSITE` [20]).

- Falls keine signifikante Homologie zu einer bereits aufgeklärten Struktur gefunden wurde, werden Ranglisten für Sequenzstrukturverwandtschaften erstellt. Dazu werden

  – verschiedene Methoden (Sequenzalignment `align` [226], Profilalignment `123D` [5], Sequenzstrukturalignment RDP [340])

  – mit verschiedenen Parametersätzen

  – auf verschiedenen Repräsentativmengen von Proteinfaltungen beziehungsweise allen in der `PDB` [29] abgelegten Proteinstrukturen angewendet.

- Konnten eine oder mehrere Faltungen als mögliche Modelle für die untersuchte Sequenz identifiziert werden, werden im nächsten Schritt die unterliegenden Sequenzstrukturalignments nach folgenden Kriterien analysiert:

  – Kompaktheit des Alignments,

  – Übereinstimmung der aus dem Modell und dem Alignment resultierenden Sekundärstruktur mit *a priori*–Informationen beziehungsweise den Ergebnissen von Sekundärstrukturvorhersagemethoden wie zum Beispiel `PHD` [288, 289] und `SOPMA` [113],

  – Analyse der Zuverlässigkeit der Alignments durch Auswertung der Pfadkontour– und Zuverlässigkeitsmatrizen [227], sofern die Alignments mit dynamischer Programmierung berechnet wurden,

  – Analyse der Pseudoenergiepotentiale zum Beispiel durch visuellen Vergleich und Korrelationsanalysen zwischen dem positionellen Energieverlauf für die native Sequenz und dem für die alinierte Sequenz,

  – Korrektheit der Topologie der Schwefelbrücken, sofern diese aus der Literatur bekannt sind oder aus der Strukturvorlage postuliert werden können,

  – Kompaktheit des aus dem Alignment resultierenden Strukturmodells; das beinhaltet zum Beispiel Untersuchungen, ob es sich um eine globuläre Struktur handelt oder ob es möglich ist, aus dem partiellen Alignmentmodell eine geschlossene Peptidkette zu bilden,

  – Zugänglichkeit von aus der Funktionsanalyse bekannten funktionellen Resten,

- Wenn eine gefundene Sequenzstrukturpaarung und das zugehörige Sequenzstrukturalignment diese Tests erfolgreich bestanden haben, wird im nächsten Schritt versucht, ein vollständiges Strukturmodell auf der Basis des Alignments zu erzeugen. Hier kommen die Standardmethoden der vergleichenden Modellierung wie das Programmpaket `MODELLER` [305] zum Einsatz.

Am Ende dieser Prozedur stehen ein oder mehrere Strukturvorschläge für die untersuchte Sequenz, zusammen mit einer Liste von Argumenten, die für oder gegen den jeweiligen Vorschlag sprechen. Falls es die durch das Alignment postulierte Ähnlichkeit zwischen der Sequenz und der Modellstruktur erlaubt, kann sogar ein vollständiges Modell konstruiert und von Experten weiter analysiert werden.

8.3 Alignmentqualität

8.3.1 Definition von Gütekriterien

Um die Qualität eines Sequenzstrukturalignments zu messen, ist es notwendig, festzulegen, was gute von schlechten Alignments unterscheidet.
Ist die dreidimensionale Struktur nur zu einem der alinierten Proteine bekannt, so ist eine objektive Bewertung eines Sequenzstrukturalignments nur schwer möglich. Eine Möglichkeit besteht darin, daß aus dem Alignment resultierende Strukturmodell zu konstruieren und dieses auf seine Plausibilität hin zu bewerten. Bei den in dieser Arbeit betrachteten Fallbeispielen handelt es sich in der Regel um Beispiele aus der Grenzzone niedriger Sequenzähnlichkeit. Das bedeutet, daß es nur selten möglich ist, vollständige Strukturmodelle mit allen Seitenketten zu bilden, die den Einsatz von Programmen wie `Procheck` [192], mit denen Proteinstrukturen auf typische Fehler und Abweichungen von Standardgeometrien getestet werden, ermöglichen würden. Außerdem wird ein so erzieltes Ergebnis stark von der Methode beeinflußt, die aus dem zu bewertenden Alignment ein vollständiges Modell generiert, so daß die Qualität des Alignments nur sehr indirekt quantifiziert werden kann. Außerdem kann ein derartiger Aufwand in wenigen Einzelfällen oder in Fällen aufgebracht werden, wo es um echte Vorhersagen geht.
Für die intensive Evaluierung einer Vorhersagemethode ist es notwendig, diese Bewertung effizient und automatisch durchführen zu können. Dies wiederum erscheint nur möglich, wenn auch die Struktur der untersuchten Sequenz aufgeklärt ist und zur Bewertung herangezogen werden kann. In diesem Fall eröffnet sich eine Reihe von Möglichkeiten zur Analyse von Alignments bezüglich ihrer Korrektheit:

- Berechnung der Struktursuperposition gemäß des Alignments mit darauf aufbauender
  - visueller Analyse,
  - Berechnung der *RMS*–Abweichung (siehe Definition 3.4) und
  - Analyse der positionellen strukturellen Abweichung in einem sogenannten *RMS*–Profil.
- Berechnung des optimalen Strukturalignments und Vergleich des berechneten Alignments gegen diesen *standard–of–truth* durch

- ein Alignment der Alignments,
- Zählung der identisch alinierten Positionen,
- einen positionellen Alignmentvergleich mittels entsprechender grafischer Aufbereitungen,
- Berechnung des mittleren Verschiebungsfehlers eines Alignments in Bezug auf ein gegebenes Referenzalignment (*mean shift error (MSE)*).

Im folgenden werden die verschiedenen Gütekriterien und Vergleichsverfahren sowie ihre Vor– und Nachteile kurz vorgestellt.
Ein in vielen Fällen sehr aussagekräftiges Maß für die Güte eines Sequenzstrukturalignments ist die *RMS*–Abweichung der gemäß des Alignments superpositionierten Strukturen (siehe Definition 3.4) zusammen mit der Anzahl der in dem Alignment alinierten Positionen. Ein Alignment ist mit Sicherheit gut, wenn ein großer Prozentsatz der beiden Strukturen mit einer niedrigen *RMS*–Abweichung (zum Beispiel ≤ 3 Å [281]) aliniert ist.
Der Umkehrschluß, daß ein Alignment in allen Teilen schlecht ist, wenn der *RMS*–Wert groß ist, gilt jedoch nicht immer. So kann das Alignment in weiten Teilen korrekt sein und durch lokale Fehler die *RMS*–Abweichung der zugehörigen Superposition groß werden (siehe zum Beispiel die Analyse des `2fd2-1fdn`–Alignments in Abschnitt 8.3.3). Außerdem zeigen Reva *et al.* in [281], daß die Wahrscheinlichkeit, in der Datenbank zufällig eine Struktur mit einer *RMS*–Abweichung ≤ 6Å zu finden, bereits bei 80 Aminosäureresten 10^{-5} und bei mehr als 160 Aminosäureresten sogar 10^{-11} beträgt. Somit kann auch ein Sequenzstrukturalignment mit einer *RMS*–Abweichung um die 6Å wertvolle Hinweise auf die Struktur der untersuchten Sequenz liefern.
Als Referenzwert kann die anhand eines Strukturalignments berechnete *RMS*–Abweichung dienen, wobei jedoch folgende Punkte zu berücksichtigen sind:

- Das Strukturalignmentproblem ist *NP*–vollständig [2], und die eingesetzten Strukturalignmentmethoden basieren auf Heuristiken (siehe Abschnitt 3.3.3), die nur Näherungen der optimalen Lösung produzieren.
- Mit verschiedenen Methoden berechnete Strukturalignments weichen zum Teil stark von einander ab. Dies kann in einigen Fällen soweit gehen, daß zwei mit unterschiedlichen Methoden berechnete Strukturalignments in keiner Position übereinstimmen [117].
- Auch unabhängig von der Berechnungsmethode ist die optimale Lösung des Strukturalignmentproblems keinesfalls eindeutig. Zu–Kang und Sippl [383] geben Beispiele für Strukturalignments bei denen trotz vergleichbarer *RMS*–Abweichung und gleicher Anzahl alinierter Reste keine Position übereinstimmt.

Beschränken sich die Fehler in einem berechneten Alignment auf lokal begrenzte Bereiche, so ist eine genauere Analyse des Alignments durch visuelle Ana-

lyse der zugehörigen Superposition oder durch Berechnung eines sogenannten *RMS*–Profiles (siehe zum Beispiel Abbildungen 8.2 und 8.3) möglich. In einem *RMS*–Profil werden die Abstände einander zugeordneter Positionen in der Superposition und zusätzlich die Abstände zu den der alinierten Position links und rechts benachbarten Positionen aufgetragen. Für ein gutes Alignment sollte die zu dem Alignment gehörende Kurve in möglichst weiten Bereichen unter den beiden anderen Kurven liegen. Weiträumige Abweichungen von dieser Regel bedeuten, daß das Alignment durch eine Verschiebung der betroffenen Sequenzsegmente verbessert werden kann.
Ein weitere Möglichkeit zur Bewertung eines Alignments besteht darin, daß berechnete Alignment gegen das *korrekte* Alignment, sozusagen gegen den *standard–of–truth*, zu vergleichen. Dieses korrekte Alignment kann zum Beispiel von einem Experten unter Einbeziehung biochemischer und phylogenetischer Zusatzinformationen erstellt worden sein. Wenn zu beiden Proteinen die dreidimensionalen Strukturen bekannt sind, wird jedoch in der Regel das Alignment aus der optimalen strukturellen Superposition abgeleitet. Dabei sollten in Grenzfällen immer die bereits diskutierten Probleme der strukturellen Superposition beachtet werden.
Für zwei gegebene Alignments (zum Beispiel ein „korrektes“ und ein berechnetes) muß das Problem des Vergleichs zweier Alignments gelöst werden. Das an der `GMD` entwickelte Programmpaket **ToPLign** [226] enthält dazu verschiedene Möglichkeiten zum Beispiel durch

- ein Alignment von Alignments,
- einen positionellen Vergleich,
- die Berechnung des mittleren Verschiebungsfehlers.

Beim Alignment von Alignments werden die zu vergleichenden Alignments als Profile betrachtet und mit einer einfachen Kostenfunktion, die die Anzahl der Identitäten maximiert, aliniert. Abbildung 5.4 zeigt das Alignment zweier Alignments mit entsprechenden Annotationszeilen. Die zur Berechnung des Alignments zweier Alignmentprofile verwendete Kostenfunktion bewertet nicht wie beim normalen Sequenzalignment den Austausch zweier Aminosäuren, sondern die Identität der jeweiligen alinierten Zweiertupel aus den Alignmentprofilen. In dieser Eigenschaft der Kostenfunktion liegt auch das Problem, mit dem diese Vergleichsmethode behaftet ist. Beim Alignment wird nur darauf geachtet, daß die in den Spalten der alinierten Profilpositionen befindlichen Aminosäuretypen übereinstimmen und nicht die hinter diesen Repräsentanten stehenden Positionen in den alinierten Sequenzen beziehungsweise Strukturen. Dieses Problem tritt umso gravierender auf je unterschiedlicher die zu vergleichenden Alignments sind. Dagegen ist es vernachlässigbar, wenn der Vergleich zweier relativ ähnlicher Alignments veranschaulicht werden soll.
Eine genauere Analyse von Alignments hat durch einen positionellen Vergleich zu erfolgen. Dies ist insbesondere dann notwendig, wenn Alignments nicht ihrer

selbst wegen oder für phylogenetische Untersuchungen verwendet werden, sondern als Grundlage für die Modellierung einer Proteinstruktur dienen sollen. Auch für den positionellen Vergleich bietet **ToPLign** entsprechende Hilfen:

- Eine Möglichkeit, zwei Alignments positionell zu vergleichen, besteht darin, zwei Alignments gemeinsam bezüglich des jeweils ersten Proteins auszurichten. Die Ausgabe erfolgt so, daß in einer Spalte die jeweils erste Zeile der beiden Alignments identisch ist (siehe Abbildung 8.5). Dazu ist es erforderlich, sogenannte Profilinsertionen einzufügen, die in Abbildung 8.5 durch das Zeichen $\sim$ von Alignmentinsertionen abgehoben sind. Die Alignments stimmen dann an einer Position überein, wenn nicht nur die ersten, sondern auch alle anderen Zeileneinträge der beteiligten Alignments identisch sind. Zur Unterstützung der visuellen Analyse werden diese Positionen in der Annotationszeile durch * hervorgehoben.
- Vergleicht man ein berechnetes Alignment mit einem Alignment, das als strukturell richtig angesehen wird, so entspricht die Anzahl der mit einem * markierten Positionen, der Anzahl der korrekt alinierten Positionen ($\#_{cor}$). Die Anzahl $\#_{cor}$ ist ein weit verbreitetes Kriterium zum Alignmentvergleich und wurde bereits im ersten Strukturvorhersagewettbewerb `CASP I` zur Bewertung der eingereichten Sequenzstrukturalignments verwendet [197]. Dieses Kriterium wird hauptsächlich dann eingesetzt, wenn die *RMS*–Abweichung groß ist, da mit ihm die richtig alinierten Bereiche des Alignments erkannt werden, die bei der *RMS*–Abweichung möglicherweise durch grobe Alignmentfehler in anderen Bereichen des Alignments überdeckt werden. Der Anteil der korrekt alinierten im Verhältnis zu allen alinierten Positionen wird auch als *Sequenzstrukturalignmentspezifizität* bezeichnet [216].
- Die oben beschriebenen Vergleichsmethoden heben nur auf die korrekt alinierten Positionen ab und lassen die anderen Positionen außer acht. In der Regel ist es für die Bewertung eines Alignments auch von Interesse, um wieviele Positionen ein Sequenzstrukturalignment bestimmte Bereiche in Bezug zum Referenzalignment verschiebt. In Abbildung 8.4 ist diese positionelle Verschiebung bezüglich des Referenzalignments durch die mehr oder weniger vertikal verlaufenden Linien repräsentiert, die die Verschiebungsfehler im Alignment visualisieren.

Der mittlere Verschiebungsfehler (*mean shift error (MSE)*) ergibt sich aus der Summe der Beträge der Verschiebungsfehler normiert auf die Anzahl der Summanden.
In der ursprünglich von Stephen Bryant zur Bewertung der Vorhersagen im `CASP II`–Wettbewerb [205] vorgeschlagenen Definition wird der Verschiebungsfehler über die Anzahl der Strukturpositionen definiert, um die eine alinierte Sequenzposition in Bezug zur gleichen Zuordnung im Referenzalignment verschoben ist.

Wie in [131] anhand eines Beispiels gezeigt wird, hat diese Definition den Nachteil, daß der so definierte mittlere Verschiebungsfehler trotz Normierung stark von der Anzahl der in den Alignments alinierten Positionen abhängt. Der Nachteil kommt insbesondere dann zum Tragen, wenn über den Verschiebungsfehler ein Vergleich unterschiedlicher Alignments in Bezug zu einem Referenzalignment durchgeführt wird. Der Nachteil entfällt, wenn der Verschiebungsfehler nicht bezogen auf die Struktur, sondern auf das Referenzalignment bezogen berechnet wird. Der Verschiebungsfehler eines Alignments ist im folgenden als die Summe der Beträge der Verschiebungen zu den im Referenzalignment alinierten Positionen definiert. In den folgenden Abschnitten wird die Qualität der mit der RDP–Methode berechneten Alignments gegen die aus verschiedenen Datenbanken (`HSSP`, `JOY`) ausgewählten, beziehungsweise mit Strukturalignmentmethoden (`SARF2`) erzeugten Alignments verglichen. Außerdem wird für die so ausgewählten Beispiele ein Vergleich mit den von anderen Sequenzstrukturalignmentmethoden berechneten Alignments durchgeführt. Als Vergleichspartner hierbei dienen:

- die an der `GMD` und am `NIH` entwickelte `123D`–Methode [5],
- das auf der doppelten dynamischen Programmierung basierende Sequenzstrukturalignment Programm von David Jones [168] in den folgenden Versionen:

  `Threader1:` Sequenzstrukturalignment auf Basis der Optimierung von Paarpotentialen.

  `Threader2:` Neuere Version, die bei der Berechnung von Sequenzstrukturalignments auch Aminosäureaustauschmatrizen und Informationen aus Sekundärstrukturvorhersagenmethoden mit einbezieht.

8.3.2 Testmengen für die Bewertung der Alignmentqualität

Sequenzstrukturalignmentmethoden, die Paarwechselwirkungen als Bestandteil der Bewertungsfunktion verwenden, basieren auf der Hypothese, daß bei ähnlichen Faltungen das Netzwerk der die Faltung stabilisierenden Wechselwirkungen konserviert ist. Zhang *et al.* zeigen, daß zwischen Proteinen mit unähnlicher Sequenz und ähnlicher Faltung bei strukturrichtigem Alignment etwa 60 bis 70% der Paarkontakte (nicht die interagierenden Aminosäuren) konserviert sind [373]. Dies setzt jedoch eine sorgfältige Auswahl der als *strukturell ähnlich* eingestuften Proteine voraus, da bei zu schwachen Kriterien für die Definition der strukturellen Ähnlichkeit die Zahl der konservierten Wechselwirkungen, insbesondere bei Proteinen niedriger Sequenzähnlichkeit, stark abnimmt und schnell unter 35% absinkt [298].
Daher hat die Auswahl der Testbeispiele für die Bewertung und den Vergleich der Alignmentqualität von Sequenzstrukturalignmentmethoden besonders sorgfältig zu erfolgen. Im folgenden werden die zur Auswahl verwendeten Kriterien

und die mit ihnen aus der HSSP und JOY–Datenbank beziehungsweise mit der Strukturalignmentmethode SARF2 ausgewählten Testbeispiele vorgestellt.
Die Verwendung der in der HSSP [310] enthaltenen Alignments erfordert zum einen das Entfernen hoch sequenzähnlicher Strukturpaare aus der Datenbasis (hier wird ein Grenzwert von maximal 80% Identität gewählt) und zum anderen sind die Alignments auch qualitativ zu filtern. Letzteres ist notwendig, da – wie schon Abbildung 3.4 zeigt – die Qualität dieser mit Methoden des multiplen Alignments erzeugten Sequenzalignments, insbesondere bei niedriger Sequenzidentität stark abnimmt und zum Test von Sequenzstrukturalignmentmethoden die gute Superpositionierbarkeit der zugehörigen Strukturen ein notwendiges Kriterium ist. Aufgrund der guten Superpositionierbarkeit ist erstens schon anhand der *RMS*–Abweichung der Struktursuperposition gemäß des berechneten Sequenzstrukturalignments entscheidbar, ob es sich um ein gutes oder schlechtes Alignment handelt, und zweitens sicher gestellt, daß ein Alignment existiert, durch das große Teile der die Struktur stabilisierenden Wechselwirkungen konserviert sind.
Zur Auswahl der zu verwendenden Alignments wurden zunächst verschiedene Repräsentativmengen der strukturaufgeklärten Proteine analysiert [131]. Diese Mengen sind der aktuellen PDBSELECT [148] entnommen und haben die Eigenschaft, daß die Sequenzidentität je zweier Repräsentanten kleiner als ein vorgegebener Schwellwert ist. Um eine nach den im folgenden genannten Auswahlkriterien angemessen große Testmenge zu erhalten, wurde die Repräsentativmenge mit einem Sequenzidentitätsgrenzwert von 35% ausgewählt (im folgenden auch als hobohm_97_35 bezeichnet). Im nächsten Schritt werden aus den dazugehörigen HSSP–Dateien die paarweisen Alignments als Referenzalignments extrahiert, die folgenden Kriterien genügen [131]:

1. beide Proteine sind strukturaufgeklärt und die in den Referenzalignments verwendeten Sequenzen stimmen mit denen der PDB–Strukturen überein,
2. es sind mindestens 70% der Reste der kürzeren Struktur aliniert,
3. die Sequenzidentität ist kleiner als 80%,
4. die *RMS*–Abweichung der Struktursuperposition gemäß des Referenzalignments ist kleiner als 3Å und
5. von zwei selektierten Alignments, bei denen die jeweils zweiten Sequenzen eine Homologie größer als 80% aufweisen, wird das Alignment mit der schlechteren Struktursuperposition aus der Testmenge entfernt.

Insgesamt gibt es in der HSSP–Datenbank 1639 Alignments eines Elements der Repräsentativmenge hobohm_97_35 mit einem zweiten ebenfalls strukturaufgeklärten Protein, doch nur 205 Paare erfüllen die beschriebenen Randbedingungen. In Tabelle 8.1 sind die 205 aus der HSSP–Datenbank verwendeten Paare zusammen mit der Identität, der Anzahl der alinierten Positionen und der *RMS*–Abweichung entsprechend des Datenbankalignments aufgeführt.

1	2	Id[%]	#	RMS[Å]	1	2	Id[%]	#	RMS[Å]	1	2	Id[%]	#	RMS[Å]
1glqB	1guhB	29.56	203	2.90	1bp2	1ppa	37.17	113	2.38	1sacE	1gnhJ	51.23	203	1.95
1nscB	1nnb	29.95	374	2.69	3grs	1ndaB	37.25	443	2.60	1ghr	1ghsB	51.49	303	1.10
1glqB	1gta	29.95	197	2.37	1mrj	1abrA	37.39	238	2.14	1frd	1fxaB	51.55	97	0.88
5nul	5fx2	30.08	133	2.74	1dsuB	2cp1	37.73	220	2.49	5timB	7timB	51.68	238	1.06
2gstB	2glrB	30.30	198	2.38	2hsp	1shfB	37.74	53	2.99	4rhv3	2hwf3	51.69	236	1.20
1glqB	1gsq	30.30	198	2.26	1csyA	1spsC	37.76	98	2.66	1qpg	1php	51.79	392	1.96
1glqB	3gstB	30.39	204	2.28	1cxc	1hrc	37.76	98	2.93	1ibeA	1pbxA	52.14	140	0.83
1nscB	2bat	30.79	380	2.65	1crb	1adl	37.80	127	1.64	1ntn	1kbaB	52.38	63	1.50
1pfxC	3rp2B	30.91	220	2.41	1ftpB	2hmb	38.17	131	1.90	1axn	1ann	53.16	316	1.72
1aab	1hma	30.99	71	2.84	1spgB	1hgcC	38.24	136	1.79	1blu	1fdn	54.00	50	2.09
1pfxC	1ppgE	31.16	215	2.55	1ftpB	1adl	38.58	127	1.86	1cxc	1c2rB	54.39	114	1.29
1lucB	1brlA	31.25	320	2.56	1bp2	1clpB	38.94	113	2.24	1dsn	1nnt	54.43	305	2.09
2gsq	1gsrB	31.31	198	2.18	1prcH	1pcrH	39.04	251	2.77	1xnb	1enxB	55.17	174	1.65
1cgo	1bbhB	31.36	118	1.65	1ars	2cstB	39.65	396	1.34	5timB	1htiB	55.27	237	1.04
1bbpD	2apd	31.45	159	2.75	2cyp	1apxD	39.83	236	2.61	1wad	2cdv	55.45	101	2.26
2fx2	4fxn	31.58	133	2.61	3minD	1mioD	40.04	447	1.62	1gadP	1ggaR	55.62	329	1.78
1eca	1flp	31.58	133	2.30	1pytC	1btp	40.27	221	2.97	1jcv	1spdB	55.63	151	1.10
1dlhE	1fruF	31.58	95	1.32	2bpa2	1gff2	40.80	174	2.11	1bndA	1bet	55.66	106	1.98
1icn	2hmb	31.78	129	1.50	1ton	1hfl	40.83	218	2.79	1frd	3fxc	55.67	97	2.66
2ak3B	3adk	31.82	176	2.82	1spgB	1pbxA	40.88	137	1.62	1cdoB	3hudB	55.76	373	2.06
1fmb	1hhp	31.91	94	2.23	1icwB	1mgsB	41.18	68	1.57	1japA	1sln	56.05	157	1.02
1hylB	2cp1	31.94	216	2.64	1aak	2uce	41.26	143	1.38	1rtp3	4pal	56.31	103	0.81
1smeB	1psaB	32.05	312	2.41	2fcr	1flv	41.32	167	2.05	1hcp	2gda	56.72	67	2.87
1smeB	1bbs	32.20	323	2.45	3rp2B	1hfl	41.36	220	2.43	1csn	1ckjB	57.09	289	1.60
1icn	1opbD	32.28	127	2.54	1abrA	1apgA	41.63	245	1.68	1fkj	1yat	57.55	106	0.77
2lhb	2pghC	32.31	130	2.17	1iscB	1mngB	41.71	187	2.83	1crb	1opbD	57.69	130	0.84
3rp2B	1ton	32.41	216	2.12	1mucB	1chrB	41.81	342	1.24	1prcL	1pcrL	58.61	273	0.75
1hmpB	1hgxB	32.56	172	1.94	1ltdB	1gox	41.96	336	2.31	1rtp3	5pal	59.26	108	0.65
1ton	3rp2B	32.86	213	2.23	1xyzB	1xas	42.23	296	2.32	1rtp3	1cdp	59.26	108	0.70
1icn	1crb	33.07	127	2.33	3ladB	1lvl	42.32	449	1.86	1plc	7pcy	60.00	95	0.83
1glqB	1hncD	33.33	198	2.13	1gpr	1gleF	42.86	154	1.73	1jcv	1srdD	60.40	149	0.80
1apa	1apgA	33.47	242	2.40	2hvm	1cnv	42.96	270	2.00	1bp2	1pshC	60.53	114	1.48
1ppfE	3rp2B	33.65	211	1.75	2fd2	1fdn	43.14	51	2.56	2cba	1huh	60.55	256	1.04
3rp2B	1ppgE	33.65	211	2.33	1dsuB	1hfl	43.18	220	2.37	1cseE	1mpt	61.34	269	1.07
1ppfE	2cp1	33.65	211	2.07	1ton	1tnl	43.18	220	1.96	1spgB	1pbxB	62.07	145	1.00
1ton	1ppgE	33.66	205	2.48	1sltB	1hlcB	43.31	127	1.01	1ccr	1hrc	62.14	103	0.55
1crb	2hmb	33.86	127	1.61	1wdcC	1mysC	43.36	143	2.27	1rtp3	3pat	62.26	106	1.22
1pfxC	1hfl	33.93	224	2.26	1mls	1myt	43.45	145	1.17	1pvc4	2hwc4	62.90	62	1.22
2mev3	2hwf3	33.94	218	2.20	1arv	1llp	43.58	335	1.29	1ltsA	1tiiA	63.74	182	1.32
1dlhE	1hsaE	34.04	94	1.26	2gstB	1gta	44.44	207	2.10	1plc	2plt	64.21	95	1.08
2sga	1p12E	34.08	179	2.11	1pmy	1pzb	44.63	121	0.97	2omf	1pho	64.83	327	1.26
1dsuB	1ppgE	34.10	217	2.93	1ntn	1ntx	44.64	56	2.80	4xiaB	6xia	65.89	384	0.84
1ftpB	1opbD	34.15	123	2.20	1bp2	1pp2R	45.13	113	2.51	1cyg	1cdg	66.17	674	0.89
1ang	1rbn	34.19	117	2.83	1spgB	2mhbB	45.52	145	1.29	1gadP	3gpdR	66.36	327	1.39
1ntn	1cvo	34.43	61	2.16	1pmlC	1kdu	46.99	83	2.40	1mhcD	1hocA	66.54	272	1.98
1edt	2ebn	34.48	261	2.87	1pfxC	1tnl	47.00	217	1.87	1plc	1plb	66.67	96	1.13
1antL	2achA	34.64	332	2.22	2mcm	1noa	47.27	110	1.29	2aaa	6taa	66.88	474	0.88
1ppfE	1tnl	34.93	209	2.16	3pga4	3ecaD	47.46	295	1.20	1ang	1agi	66.95	118	0.86
1pfxC	2cp1	34.98	223	2.19	2acq	1ral	47.67	300	1.34	1vls	2asr	68.09	141	2.26
1ntn	1cdtB	35.00	60	2.45	2achA	9apiA	47.75	333	1.28	451c	1cch	68.29	82	1.79
1avdB	1srjB	35.14	111	2.19	1hngB	1hnf	48.00	175	2.35	9pap	1ppo	68.40	212	1.23
1ppfE	1hfl	35.35	215	1.93	1xnb	1xyn	48.30	176	2.02	1mhcD	1hsaD	68.98	274	1.76
1ofv	2fcr	35.71	168	2.59	1frd	1fxiD	48.42	95	1.93	1ofv	1flv	70.66	167	0.62
1hip	1hpi	35.71	70	2.44	2hpeB	1hhp	48.48	99	1.37	1prtH	1prtI	70.92	196	1.06
1dsuB	1ton	35.94	217	2.66	1ccr	1cry	48.48	99	0.99	6fabL	1reiB	71.03	107	1.02
1icwB	1rhpD	35.94	64	1.66	1ino	2prd	48.82	170	1.53	1fbaD	1ald	71.10	353	2.22
1ton	2cp1	36.07	219	2.16	1ivd	1nnb	48.83	383	1.28	2mev2	1tme2	72.29	249	1.43
1cxc	1cry	36.08	97	2.82	1tys	1tis	49.03	259	1.96	1gpl	1hplB	72.33	430	1.56
1poh	1ptf	36.14	83	1.16	2ihl	2eql	49.22	128	1.57	1lnh	1yge	72.72	832	1.71
1ptf	1hdn	36.14	83	1.62	3rp2B	2cp1	49.33	223	0.74	8tlnE	1npc	73.10	316	0.86
1ldm	1ldnH	36.24	298	1.88	1axn	1ain	49.68	314	1.50	1hsbB	1fruF	74.75	99	0.67
3rp2B	1tnl	36.28	215	1.89	1nhkR	1ndlC	50.00	138	0.91	1apa	1pagB	75.38	260	0.86
2mcm	1akp	36.36	110	2.80	1axn	1anxC	50.16	315	1.89	1gpl	1lpbB	76.28	430	1.02
1abrA	1mrk	36.40	239	1.90	1prcM	1pcrM	50.17	303	1.47	1ilb	2mib	78.67	150	1.24
1allA	1cpcB	36.54	156	2.27	1hrjB	1hunB	50.77	65	2.17	1plc	9pcy	78.79	99	0.77
1pfxC	1ton	36.65	221	2.64	1rtp3	1rro	50.93	108	0.80	1burV	1rldT	78.86	123	0.78
1dokB	1hunB	36.76	68	1.95	2cae	8catB	51.06	470	1.14	2lbp	2liv	79.36	344	0.88
2fcr	1ofv	36.90	168	2.36	1frd	1frrB	51.09	92	1.76					
1dsuB	1tnl	37.16	218	1.72	1meyG	1zaaC	51.22	82	0.93					

Tabelle 8.1: Aus der HSSP–Datenbank verwendete paarweise Alignments (205): Identität (Id), Anzahl (#) alinierter Positionen und RMS–Abweichung (RMS) (sortiert nach Id) [131].

Die JOY–Datenbank [306] enthält eine handverlesene Sammlung von Strukturalignments, die mit den Programmen COMPARER [304] beziehungsweise dem multiplen Strukturalignmentverfahren MNYFIT [332] erzeugt und teilweise von Hand nachgebessert wurden. In Tabelle 8.2 sind die 59 aus der JOY–Datenbank mit den beschriebenen Kriterien ausgewählten Paare zusammen mit der Identität, der

Anzahl alinierter Positionen und der *RMS*–Abweichung entsprechend des Datenbankalignments aufgeführt. Die JOY–Alignments stellen einen schwierigeren Test

1	2	Id[%]	#	RMS[Å]	1	2	Id[%]	#	RMS[Å]	1	2	Id[%]	#	RMS[Å]
2mm1	1lh1	17.01	147	2.84	6ldh	1ldb	35.17	290	1.50	1ads	1ral	47.67	300	1.24
3fxn	2fcr	20.15	134	2.61	1hhl	1alc	35.25	122	1.84	1st3	1thm	47.74	266	1.35
2mm1	2lhb	21.17	137	1.97	2alp	2sga	35.75	179	1.51	9pap	2act	48.34	211	0.98
2mm1	2hbg	21.43	140	2.18	2mcm	1acx	36.11	108	1.63	1tis	3tms	49.03	259	2.36
2cpp	1cpt	21.73	382	2.83	1bbs	4cms	38.29	316	1.67	5pal	5cpv	50.00	108	0.93
2mm1	1ecd	22.06	136	1.83	3est	2ptn	38.64	220	1.33	1hhl	1lz1	56.59	129	1.07
2mm1	1mba	22.38	143	2.17	1bp2	1ppa	38.94	113	1.58	1fkb	1yat	57.01	107	0.85
3fxn	1ofv	23.88	134	2.55	3est	2gch	39.56	225	1.54	1fus	1rds	57.84	102	1.46
3fxn	1flv	24.06	133	2.48	6ldh	1llc	39.62	313	2.26	1fus	1rnt	59.80	102	1.33
1bbs	3app	24.60	309	2.47	1bbs	5pep	39.63	323	1.57	1st3	1sbt	59.85	269	1.22
1bbs	4ape	25.48	314	2.96	3aat	1ama	39.90	396	1.55	1st3	1sbc	60.97	269	1.06
1bbs	2apr	26.20	313	2.25	1bp2	1bbc	41.74	115	1.68	1ca2	2cab	61.57	255	1.04
3dfr	8dfr	28.12	160	2.17	1f3g	1gpr	42.28	149	1.51	6ldh	2ldx	64.02	328	1.80
3fxn	1fx1	29.20	137	1.97	1cry	1ycc	43.14	102	1.61	1pho	1omf	64.53	327	0.90
1bbs	1mpp	29.81	322	2.97	1cry	1ccr	44.12	102	1.65	9pap	1ppo	68.40	212	1.03
3est	1ton	30.77	221	2.22	5pal	1omd	44.86	107	0.78	1cor	351c	69.14	81	1.56
8tlnE	1ezm	31.07	280	2.87	2mm1	1myt	45.21	146	1.31	6ldh	5ldh	71.43	329	2.59
3est	1sgt	34.11	214	2.07	1tis	4tms	46.79	265	2.65	8tlnE	1npc	72.15	316	1.00
1hc1	1lla	34.19	582	2.28	2mcm	1noa	46.85	111	1.21	2lbp	2liv	79.07	344	0.78
1cry	2c2c	34.34	99	1.63	5pal	1pal	47.66	107	1.03					

Tabelle 8.2: Aus der JOY–Datenbank verwendete paarweise Alignments (59): Identität (Id), Anzahl (#) alinierter Positionen und *RMS*–Abweichung (RMS) (sortiert nach Id) [131].

für die Sequenzstrukturalignmentmethoden da, weil es sich bei ihnen um auf struktureller Information abgeleitete Alignments handelt, die nicht ohne weiteres mit Sequenzalignmentmethoden reproduziert werden können. Im Vergleich zu den Beispielen aus der HSSP–Datenbank liegt die Sequenzidentität der Referenzalignments aus JOY etwas niedriger und beginnt bereits bei 17% für die zwei Globine 2mm1 und 1lh1. Daher ist eigentlich zu erwarten, daß auf dieser Testmenge Methoden, die auch Sequenzinformation verwenden, schlechtere Ergebnisse als auf der HSSP–Menge liefern.
Die weitaus schwierigste Herausforderung für Sequenzstrukturalignmentmethoden wie RDP und 123D ist der Vergleich gegen Struktursuperpositionierungsverfahren, wie zum Beispiel SARF2 [7, 8]. SARF2 ist ein Strukturalignmentprogramm, welches zunächst nach gut superpositionierbaren Sekundärstrukturelementen sucht und dann darauf aufbauend versucht, die Superposition auf weitere Reste auszudehnen. Um repräsentative Aussagen zu ermöglichen, wurden für diesen Test die 251 Proteine der Repräsentativmenge hobohm_96_25 [148] ausgewählt, die nach der SCOP–Klassifizierung [243] aus nur einer Domäne bestehen. Die 251 Strukturen wurden mit SARF2 paarweise superpositioniert und die aus der Superpositionierung abgeleiteten Alignments den bereits beschriebenen Kriterien unterworfen.
In Tabelle 8.3 sind die 73 mittels SARF2 bestimmten Paare zusammen mit der Sequenzidentität, der Anzahl alinierter Positionen und der *RMS*–Abweichung entsprechend des SARF2–Strukturalignments aufgeführt. Die Sequenzidentität gemäß der Strukturalignments liegt zwischen 6.5% für die Ferredoxin ähnlichen Faltungen 1aps und 2bopA und 28.9% für die Ribonukleasen 7rsa und 1onc. Für die Ribonukleasen wird damit sogar der bei der Ableitung der Repräsentativmenge verwendete Grenzwert von 25% Sequenzidentität überschritten. Dies liegt daran,

1	2	Id[%]	#	RMS[Å]	1	2	Id[%]	#	RMS[Å]	1	2	Id[%]	#	RMS[Å]
1aps	2bopA	6.45	62	2.73	2hbg	1ash	14.84	128	2.22	1dynA	1pls	20.88	91	2.20
1was	256bA	7.79	77	2.72	1hrm	2gdm	15.15	132	2.04	1epaB	1mup	21.28	141	1.98
2bopA	1aps	8.06	62	2.87	2fal	1pbxA	15.56	135	2.19	1mup	1epaB	21.28	141	1.98
1eca	1pbxA	8.20	122	2.38	1arv	2cyp	15.81	234	2.02	1eca	1hlb	22.05	127	2.02
7pcy	2azaA	10.39	77	2.57	3cla	1dpb	16.67	156	2.43	1bbt1	2mev1	22.29	166	1.94
1bgc	1rcb	10.99	91	2.16	1lkkA	2pleA	16.88	77	2.66	1hrm	1pbxA	22.30	139	1.73
2fal	1ash	11.43	140	1.83	256bA	2ccyA	17.05	88	1.73	1rcf	4fxn	23.13	134	2.08
1bgc	1lki	12.30	122	2.41	1arb	1try	17.18	163	2.04	4fxn	1rcf	23.44	128	2.30
1bbpA	1hbq	12.31	130	2.04	1hrm	1hlb	17.19	128	1.85	2pleA	1lkkA	23.53	85	2.10
1eca	2gdm	12.40	129	2.78	1dhr	1hdcA	17.53	194	2.18	1pyp	2prd	23.60	161	1.87
1ash	1hlb	12.69	134	2.30	2fal	2gdm	17.60	125	2.80	2fal	1eca	23.66	131	2.05
4fgf	1hce	12.73	110	2.24	1pbxA	1hlb	17.60	125	2.21	3dfr	1dyr	23.87	155	1.57
1hrm	1ash	12.78	133	1.91	1hdcA	1dhr	17.62	193	2.20	2fal	1hrm	23.94	142	2.05
1ltsD	1prtF	13.58	81	2.09	2hbg	2gdm	18.11	127	2.29	1pne	2acg	23.97	121	1.66
1pbxA	1ash	13.71	124	2.15	2gdm	1hlb	18.32	131	2.67	2acg	1pne	23.97	121	1.66
4fgf	1il b	13.76	109	2.22	1il b	1hce	18.63	102	2.34	1cpcA	1cpcB	24.18	153	1.64
1il b	1wbc	13.91	115	2.32	2fal	1hlb	18.66	134	2.31	1dyr	3dfr	24.52	155	1.77
3chy	5p21	13.98	93	2.71	1hrm	1eca	19.12	136	1.70	1itg	1vsd	25.23	111	1.71
1prtD	1prtF	14.10	78	1.93	2hbg	1pbxA	19.55	133	1.56	1vsd	1itg	25.23	111	1.71
1eca	1ash	14.29	133	1.97	1epaB	1hbq	19.70	132	2.08	1bbt2	1pvc2	27.17	184	1.34
1minA	1minB	14.29	357	2.53	2hbg	1hlb	20.00	125	2.23	1bbt3	4rhv3	28.02	207	1.67
1minB	1minA	14.41	354	2.50	2hbg	2fal	20.00	135	2.30	1onc	7rsa	28.89	90	1.46
2gdm	1pbxA	14.53	117	2.33	2hbg	1eca	20.16	124	2.49	7rsa	1onc	28.89	90	1.46
1hbq	1bbpA	14.73	129	1.99	1hbq	1epaB	20.45	132	1.95					
256bA	2hmzA	14.81	81	2.74	2hbg	1hrm	20.77	130	1.97					

Tabelle 8.3: Mit SARF2 ermittelte Paare (73): Identität (Id), Anzahl (#) alinierter Positionen und RMS–Abweichung (RMS) (sortiert nach Id) [131].

daß zur Berechnung der Repräsentativmenge Sequenzalignmentmethoden verwendet werden und diese bei den üblicherweise verwendeten Parametrisierungen mehr Positionen, zum Beispiel sequenzunähnliche Schleifenbereiche, in das Alignment einbeziehen als dies durch Strukturalignmentverfahren wie SARF2 geschieht, die strukturell unterschiedliche und damit häufig auch sequenzunähnliche Schleifen *per definitionem* nicht alinieren.
Um den im Threading üblichen Fall zu simulieren, daß die Struktur des untersuchten Proteins nicht bekannt ist, wurde die jeweils erste Struktur eines Paares als bekannt und die andere als unbekannt angenommen, so daß es sich bei den im folgenden durchgeführten Tests um Sequenzstrukturalignments handelt, bei denen zwar beide Strukturen bekannt sind, aber eine der Strukturen ignoriert wird. Die Tabelle 8.4 zeigt die Anzahl der so generierten Testbeispiele und gibt eine

	#	Alignment–länge	RMS [Å]	Identität [%]	alinierte [%] Positionen	# CATH (siehe Abschnitt 3.3.5)
HSSP	205	50 – 832	0.6 – 3.0	30 – 80	90 – 100	3 *unclassified*, 38 α, 73 β, 60 $\alpha\beta$, 1 *few sec. struc.*, 30 *multi-domain*
JOY	59	81 – 532	0.8 – 3.0	17 – 80	91 – 100	16 α, 15 β, 21 $\alpha\beta$, 7 *multi-domain*
SARF2	73	120 – 661	1.3 – 2.9	6.5 – 29	70 – 100	34 α, 18 β, 21 $\alpha\beta$

Tabelle 8.4: Eigenschaften der aus HSSP [310], JOY [306] und mit SARF2 [7] ausgewählten Testmengen [131].

Übersicht über die aus HSSP, JOY beziehungsweise mit dem Strukturalignmentprogramm SARF2 ausgewählten Alignments und die zugehörigen Strukturen. Die

Sequenzidentität kann auch für Sequenzstrukturalignmentmethoden die paarweise Wechselwirkungen in die Alignmentberechnung einbeziehen als ein Kriterium für die Schwierigkeit der Berechnung eines qualitativ guten Alignments angesehen werden. Dies liegt zum einen daran, daß das Wechselwirkungsnetzwerk desto stärker zwischen zwei Proteinen konserviert ist, je höher die Sequenzidentität ist, und zum anderen daran, daß auch Wechselwirkungspotentiale in der Regel Wechselwirkungen bevorzugen, bei denen die beteiligten Aminosäuren konserviert sind. Die fünfte Spalte der Tabelle 8.4 belegt, daß die Schwierigkeit der Beispiele von den Testbeispielen aus der `HSSP` über die aus `JOY` extrahierten Beispiele bis hin zu den reinen Strukturalignments zunimmt, die mit `SARF2` berechnet wurden. Damit kann anhand dieser Testbeispiele gleichzeitig eine Abschätzung versucht werden, bis zu welchem Schwierigkeitsgrad Sequenzstrukturalignmentmethoden anwendbar sind, um für die vergleichende Modellierung verwendbare Sequenzstrukturalignments zu berechnen.
Durch die restriktiven Auswahlkriterien, was die *RMS*–Abweichung und den Anteil der superpositionierbaren Reste betrifft, soll sichergestellt werden, daß ein Alignment existiert, durch welches das die Faltungen stabilisierende Wechselwirkungsnetzwerk (nicht die unbedingt die daran beteiligten Aminosäuren) konserviert wird. Inwieweit und insbesondere bis zu welchem Detailgrad der Unterscheidung von Wechselwirkungen dies der Fall ist, ist zu untersuchen, soll aber hier ein Desiderat bleiben.
Durch die Existenz eines derartigen Alignments ist sichergestellt, daß Methoden wie `Threader` und RDP aufgrund ihrer verwendeten Bewertungsfunktionen die Chance haben, ein strukturrichtiges Alignment mit guter Bewertung zu finden.

8.3.3 Vergleich auf der Basis von multiplen `HSSP`–Alignments

Die Abbildungen 8.1 und 8.6 vergleichen die Güte der mit der RDP–Methode für die aus der `HSSP`–Datenbank ausgewählten Sequenzstrukturpaarungen berechneten Alignments mit denen, die mit `123D` [5] beziehungsweise den Programmen `Threader1` [168] und `Threader2` [167] berechnet wurden. Da es sich bei `123D` um eine sehr schnelle Methode handelt, wurden die Vergleichsalignments für 32 plausible Parametersätze berechnet und anschließend derjenige `123D`–Parametersatz für den Vergleich mit der RDP–Methode ausgewählt, der die besten Ergebnisse für die Testmenge gemäß des hier verwendeten *RMS*–Kriteriums liefert. Für die Testläufe mit dem Programmen `Threader1` und `Threader2` wurden die Standardparameter verwendet, die von David T. Jones mit dem Programm mitgeliefert werden [167]. Für die RDP–Methode werden die Parameter und Potentiale verwendet, die sich in Abschnitt 7.7 als am besten herausgestellt haben.
In der Abbildung 8.1 sind die Alignments nach ihrer Güte bezüglich des *RMS*–Kriteriums als Histogramm aufgetragen. Die dunkelblauen Balken zeigen die *RMS*–Abweichung für die aus `HSSP` ausgewählten Referenzalignments. Gemäß eines der Kriterien für die Auswahl eines Alignments liegt die *RMS*–Abweichung

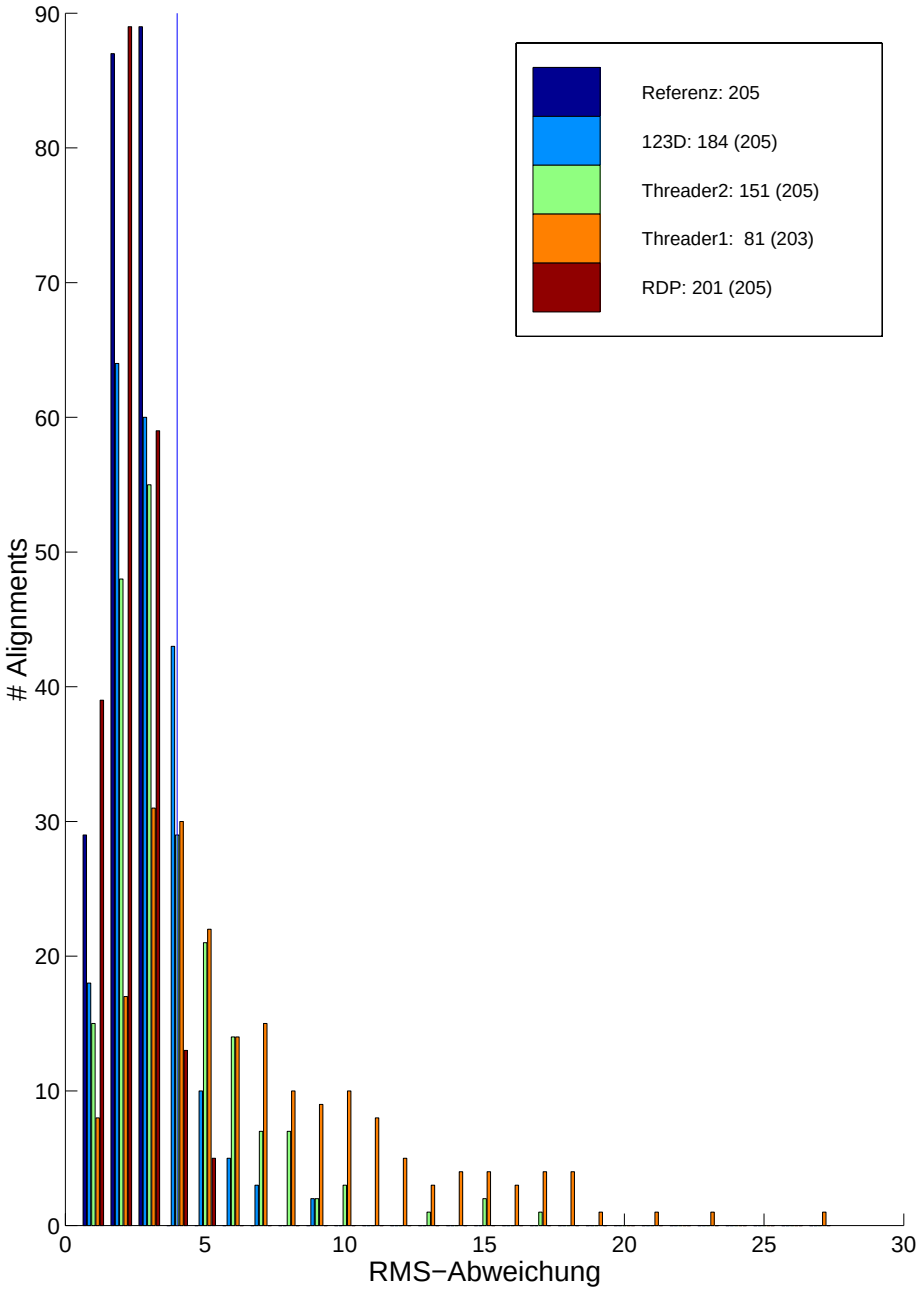


Abbildung 8.1: Histogramm der *RMS*–Abweichungen der von 123D (hellblau), `Threader2` (grün), `Threader1` (orange) und RDP (rot) berechneten Sequenzstrukturalignments im Vergleich für 205 `HSSP` (dunkelblau) Alignments. Die Legende gibt für jede Methode die Anzahl der Alignments (gesamt) an, deren *RMS*–Abweichung kleiner als die des `HSSP`–Alignments + 2Å ist.

aller 205 Referenzalignments zwischen 0 und 3Å, wobei die meisten Alignments eine *RMS*–Abweichung von etwa 2Å haben. Es muß hier nochmals betont werden,

daß diese Alignments bei weitem keine repräsentative Auswahl aus der HSSP–Datenbank darstellen, was ihre Qualität bezüglich des *RMS*–Kriteriums betrifft. Der qualitative Unterschied zwischen den verglichenen Sequenzstrukturalignmentmethoden zeigt sich bereits auf den ersten Blick, wenn man die Anzahl der Alignments betrachtet, für die die *RMS*–Abweichung der zugehörigen Strukturüberposition die des Referenzalignments maximal um 2Å überschreitet.
Die RDP–Methode erreicht hier mit 201 von 205 Alignments einen Wert, der von den zum Vergleich herangezogenen Methoden unerreicht bleibt (siehe Legende von Abbildung 8.1). Abgeschlagen mit 184 von 205 *guten* Alignments folgt die Methode 123D. Dabei ist anzumerken, daß zum Vergleich die 123D–Parameterkombination herangezogen wurde, die von 32 weiteren Parametersätzen für diese Testmenge bezüglich des verwendeten Kriteriums die besten Ergebnisse lieferte. Das verbreitet eingesetzte Programm Threader berechnet in diesem Test in seiner aktuellen Version [167] für 151 Fälle brauchbare Alignments und in seiner älteren Version [168] nur 81 Sequenzstrukturalignments, die dem Kriterium genügen.
Die mit 123D für diese Beispiele berechneten Alignments sind in den hellblauen Balken zusammengefaßt. In 184 Fällen überschreitet das 123D–Alignment die durch die Referenz vorgegebene *RMS*–Abweichung um weniger als 2Å. In 81 Fällen ist die *RMS*–Abweichung für das 123D–Alignment deutlich größer als die des Referenzalignments. In 20 Fällen überschreitet die *RMS*–Abweichung für das 123D–Alignment die zur Orientierung eingezeichnete 4Å–Marke.
Die *RMS*–Abweichung für die 123D–Alignments 2ak3B-3adk und 1ton-1hf1 beträgt 8.0Å beziehungsweise 8.4Å. Im folgenden werden die Ursachen für derartig hohe Abweichungen analysiert. Abbildung 8.2 zeigt in der roten Kurve den posi-

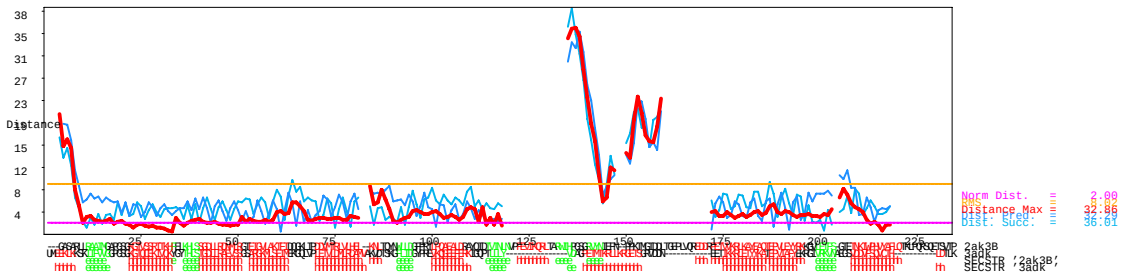


Abbildung 8.2: Analyse des positionellen Abstandes in der optimalen Superposition gemäß des 123D–Alignments der Sequenz von 3adk gegen die Struktur von 2ak3B (*RMS* 8.0Å).

tionellen Abstand zweier von 123D alinierter Aminosäurereste in der zugehörigen optimalen Superposition für die Adenylatkinasen 3adk und 2ak3B. Es zeigt sich, daß neben vernachlässigbaren Fehlern am Anfang des Alignments der Hauptfehler dieses Alignments darin besteht, daß im mittleren Bereich eine Helix von 2ak3B auf einen Strang von 3adk gelegt wird.
Ein Vergleich mit dem in Abbildung 8.3 dargestellten positionellen Abstand

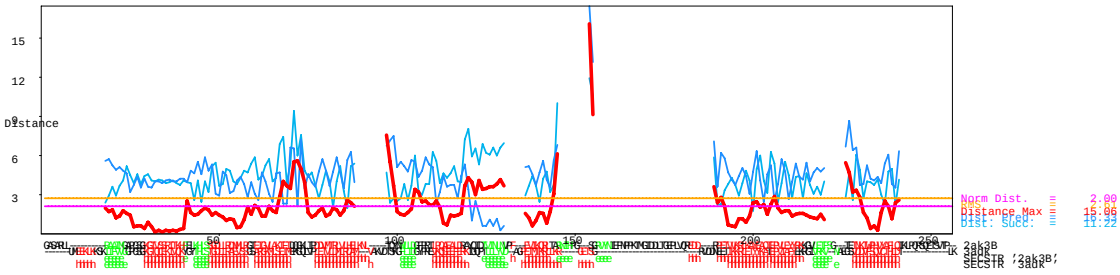


Abbildung 8.3: Analyse des positionellen Abstandes in der optimalen Superposition gemäß des RDP–Alignments der Sequenz von `3adk` gegen die Struktur von `2ak3B` (*RMS* 2.6Å).

gemäß des RDP–Alignments zeigt, wieso die RDP–Methode für die Adenylatkinase `3adk` (195 Aminosäuren) ein Sequenzstrukturalignment mit der Adenylatkinase `2ak3B` (226 Aminosäuren) mit einer *RMS*–Abweichung von 2.6Å bei 165 alinierten Positionen und 27.7% Identitäten errechnet. Der wesentliche Unterschied besteht in der richtigen Zuordnung der in dem 123D–Alignment fehlerhaft alinierten Helix.

Auch für die zwei Serinproteasen `1ton` (235) und `1hf1` (234) berechnet die RDP–Methode ein Alignment mit einer *RMS*–Abweichung von nur 2.5Å bei 193 alinierten Positionen und einer Sequenzidentität von 41.6%. In der Abbildung 8.4 werden die von `Threader1`, `Threader2`, `123D` und RDP berechneten Alignments mit dem Referenzalignment aus der `HSSP`–Datenbank verglichen. Die Linien geben neben der Information, welche Positionen im Referenzalignment aliniert sind, auch die Verschiebung bezüglich des Referenzalignments an. Das RDP–Alignment (2.5Å, 193) unterscheidet sich vom Referenzalignment (2.8Å, 218) im wesentlichen durch eine geringere Anzahl (25) alinierter Positionen. Die Abbildung 8.4 zeigt deutlich, daß neben weiteren kleinen Abweichungen der wesentliche Fehler des `123D`–Alignments im zentralen Bereich des Alignments liegt, wo eine notwendige Einfügung an der falschen Stelle vorgenommen wird.

Desweiteren zeigt Abbildung 8.4 ein typisches Beispiel für den Fortschritt zwischen den Versionen `1` [168] und `2` [167] der vermarkteten Threadingsoftware `Threader`. Während `Threader1` für die Serinproteasen `1ton` und `1hf1` noch ein Sequenzstrukturalignment mit 16.7Å *RMS*–Abweichung berechnet, liefert die aktuelle Version `Threader2` ein Alignment mit 3.5Å *RMS*–Abweichung, in dem aber im Vergleich zum RDP–Alignment (*RMS* 2.5Å) bezüglich des `HSSP`–Alignments mit 2.8Å *RMS*–Abweichung immer noch einige Reste falsch zugeordnet sind.

Mit dem bereits vermarkteten Programm `Threader1` [168] konnte für zwei der Beispiele kein Alignment berechnet werden, da der Sekundärstrukturanteil der beiden Strukturen von dem Programm als zu gering angesehen wurde. Für die verbleibenden 203 Beispiele wurde nur für 81 Fälle von `Threader1` Sequenzstrukturalignments berechnet, deren *RMS*–Abweichung die des Referenzalignments um weniger als 2Å überschreitet. Für 40 (15) der von `Threader1` berechneten Align-

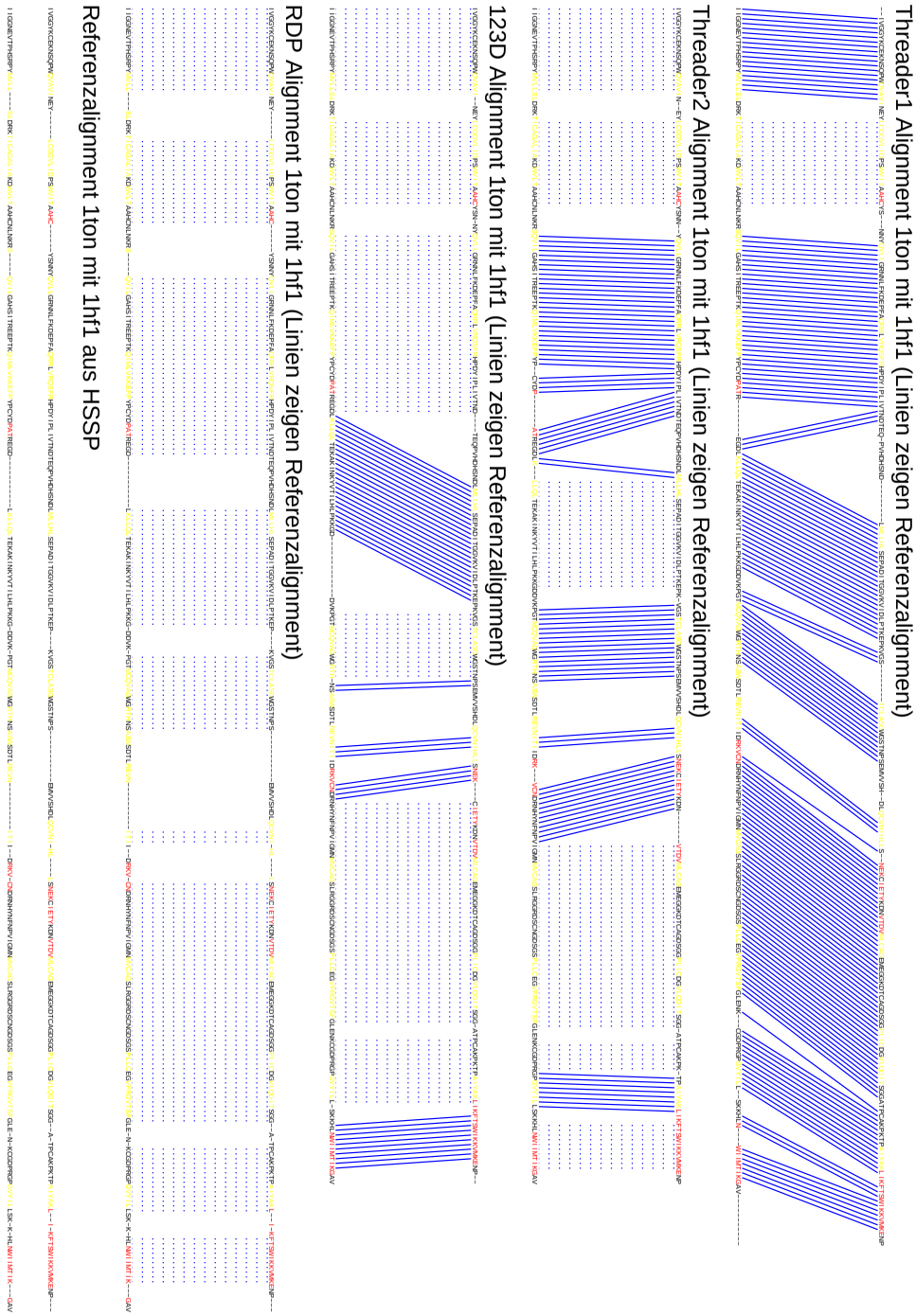


Abbildung 8.4: Positioneller Vergleich der Alignments zweier Serinproteasen (`1ton`, `1hf1`) berechnet mit `Threader1` (oben, *RMS* 16.7Å), `Threader2` (*RMS* 3.5Å), `123D` (*RMS* 8.4Å) und RDP (*RMS* 2.5Å) in Bezug zum `HSSP`–Referenzalignment (unten, *RMS* 2.8Å). Die Linien zeigen die Verschiebung bezüglich des Referenzalignments.

ments liegt die *RMS*–Abweichung über 10 (15)Å und in 3 Fällen sogar über 20Å. Zumindest in diesen 40 von 203 Fällen kann davon ausgegangen werden, daß das von `Threader1` berechnete Alignment vollkommen falsch ist. Auf eine detaillierte Analyse der `Threader1`–Ergebnisse kann verzichtet werden.
Allgemein bestätigt sich die auch zu erwartende Verbesserung von `Threader2` gegenüber `Threader1`, die sich bereits am Einzelbeispiel aus Abbildung 8.4 andeutet. Sowohl der allgemeine Vergleich (siehe Abbildung 8.1) als auch der paarbezogene Vergleich (siehe Abbildung 8.6) belegen den Fortschritt. Für 151 `Threader2`–Alignments liegen die *RMS*–Abweichungen weniger als 2Å über den durch die Referenzalignments vorgegebenen Richtwerten. Dies bedeutet eine Verdopplung der Anzahl der guten Alignments gegenüber `Threader1` (siehe oben).
Wie Abbildung 8.1 zeigt, überschreiten die *RMS* Abweichungen der mit der Version `Threader2` berechneten Alignments jedoch immer noch in 22 von 205 Fällen teilweise deutlich die 6Å–Marke. Zum Vergleich überschreiten nur 5 RDP–Alignments für diesen Test überhaupt die 4Å–Marke (siehe Abbildung 8.1).

```
Theader 1cxc_000:QEG~~~~~GDPEAGAKAFN~QCQTCHVIVDDSGTTIAGRNAKTGPNLYGV---VGRTAGT
Theader 1hrc_000:------------------------------------UGDVEKGKKIFVQKCAQCHTVEK
Allowed_000:       ..................................................
HSSP    1cxc_000:~~~QE-G-GDPEAGAKAFN-QCQTCHVIVDDSGTTIAGRNAKTGPNLYGV~~~VGRTAGT
HSSP    1hrc_000:~~~~~UGUGDVEKGKKIFVQKCAQCHTVEK-------GGKHKTGPNLHGL~~~FGRKTGQ

Theader 1cxc_060:QADFKGYGEGMKEAGAKGL-----AWDEEHFVQYVQDPTKFLKEYTGDAKAKGKMTF---
Theader 1hrc_060:GGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMEYLENPKKYIPGTKMIF
Allowed      060:............................................................
HSSP 1cxc   _060:QADFKGYGEGMKEAGAKGL~~~~~AWDEEHFVQYVQDPTKFLKEYTGDAKAKGKMTF~~~
HSSP 1hrc   _060:APGFTYTDANKN----KGI~~~~~TWKEETLMEYLENPKKYIPGT--------KMIF~~~

Theader 1cxc_120:-KLKKEADAHNIWAYLQQVAVRP~~~~~~~~~
Theader 1hrc_120:AGIKKKTEREDLIAYLKKATNE~~~~~~~~~~
Allowed      120:******************
HSSP    1cxc_120:-KLKKEADAHNIWAYLQQ~~~~~VAVRP----
HSSP    1hrc_120:AGIKKKTEREDLIAYLKK~~~~~~~~~~ATNE
```

Abbildung 8.5: Positioneller Alignmentvergleich (siehe Abschnitt 8.3.1) des `Threader2`–Alignments für die Zytochrom–C–Proteine `1cxc` und `1hrc` mit dem HSSP–Alignment (Die 18 identisch alinierte Positionen sind mit Sternen annotiert).

In den Fällen, wo die *RMS*–Abweichung größer als 6Å ist, stimmt das `Threader2`–Alignment, wie zum Beispiel für das Paar `1cxc` und `1hrc` in Abbildung 8.5 gezeigt wird, nur an sehr wenigen Stellen (im Beispiel 18) mit dem Referenzalignment überein. Damit bestätigt das Maß der korrekt alinierten Positionen die *RMS*–Abweichung (für das Beispiel 12.8Å).
Insgesamt zeigt sich schon hier, daß das Programm `Threader2`, trotz signifikanter Verbesserungen gegenüber `Threader1`, auf dem `HSSP`–Datensatz schlechtere Alignments berechnet als die an der GMD entwickelten Methoden `123D` und RDP. Noch deutlicher wird dies, wenn wie in Abbildung 8.6 die *RMS*–Abweichungen für die einzelnen Alignments aufgetragen werden.
In Abbildung 8.6 sind die einzelnen Alignments nach der *RMS*–Abweichung des jeweiligen RDP–Alignments aufsteigend sortiert. Das RDP–Alignment erreicht

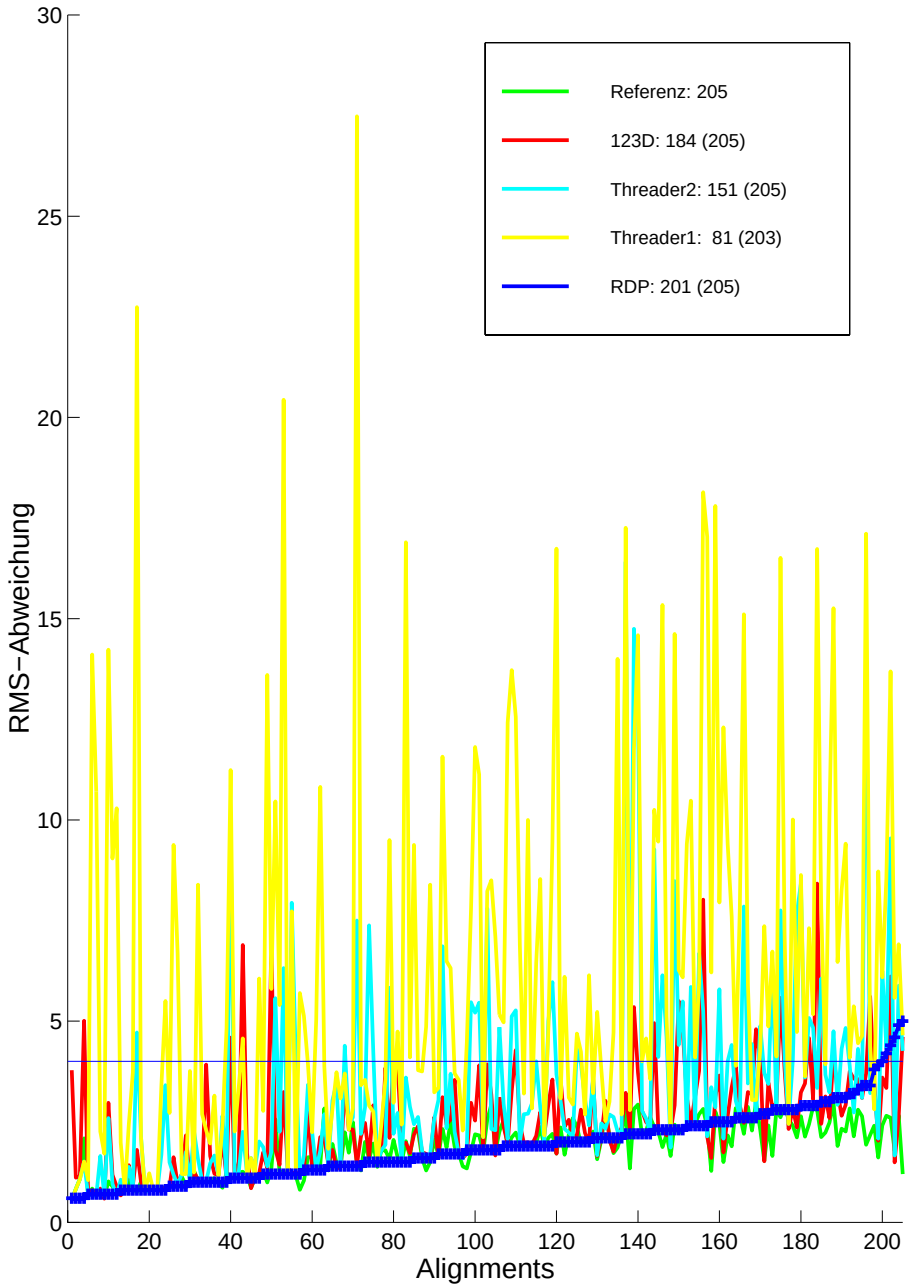


Abbildung 8.6: Vergleich der *RMS*–Abweichungen der von `123D` (rot), `Threader2` (hellblau), `Threader1` (gelb) und RDP (blau) berechneten Sequenzstrukturalignments für 205 `HSSP` (grün) Alignments (aufsteigend sortiert nach RDP).

in sehr vielen Fällen die Qualität des `HSSP`–Referenzalignments (grüne Kurve). Die Werte für die `123D`–Alignments (rote Kurve) liegen im Durchschnitt deutlich über denen der RDP–Alignments. Damit bestätigt sich der anhand der zuvor betrachteten Einzelbeispiele abzeichnende Trend, daß die RDP–Methode bessere

Alignments liefert als das schnelle 123D und somit einen signifikanten Beitrag zur Alignmentqualität gemessen an dem *RMS*–Kriterium leistet. Dieser Trend wird auch maßgeblich durch die Tatsache gestützt, daß nur 5 RDP–Alignments für diesen Test die 4Å–Marke überschreiten, während für 123D in 21 Beispielen Sequenzstrukturalignments mit einer *RMS*–Abweichung größer als 4Å und für Threader1 und Threader2 noch deutlich mehr Alignments mit großen *RMS*–Abweichungen erwartet werden müssen.
Die 205 mit Threader2 berechneten Alignments sind in Abbildung 8.6 als hellblaue Kurve aufgetragen. Bis auf wenige Ausnahmen liegt der *RMS* der Threader2–Alignments immer über dem der RDP–Alignments und in aller Regel auch über dem der 123D–Alignments. Die 203 mit Threader1 berechneten Alignments sind in Abbildung 8.6 als gelbe Kurve gezeigt. Die Threader1–Alignments sind in fast allen Fällen schlechter als die Threader2–Alignments und damit fast immer wesentlich schlechter als die mit der RDP–Methode berechneten Alignments.
Von den RDP–Alignments haben nur 5 eine *RMS*–Abweichung größer als 4Å und keines eine größere als 5Å. Dabei handelt es sich bis auf eine Ausnahme um Proteine mit mehr als 300 Aminosäuren, für die auch ein Alignment zwischen 4 und 5Å nicht direkt als schlecht einzustufen ist. Die einzige Ausnahme bildet das Alignment der Sequenz der Serinproteasen 1ton (227 Reste) gegen die Struktur der Serinprotease 1dsuB (228 Reste) mit einer *RMS*–Abweichung von 4.9Å, während das Referenzalignment nur eine *RMS*–Abweichung von 2.7Å aufweist. Abbildung 8.7 vergleicht die mit den verschiedenen Methoden für die Proteine 1dsuB und 1ton berechneten Sequenzstrukturalignments. Das kompaktere 123D–Alignment ist in diesem Fall nicht nur bezüglich der *RMS*–Abweichung, sondern auch bezüglich des Verschiebungsfehlers klar besser als das RDP–Alignment.
Abbildung 8.6 zeigt, daß es sich dabei um eine der wenigen Ausnahmen auf dieser Testmenge handelt, wo das RDP–Alignment deutlich schlechter als ein mit einer anderen Methode berechnetes Alignment ist.

8.3.4 Vergleich auf der Basis von JOY–Strukturalignments

Die Abbildungen 8.8 und 8.9 vergleichen die Güte der mit der RDP–Methode für die aus der JOY–Datenbank ausgewählten Sequenzstrukturpaare berechneten Alignments mit denen, die mit 123D [5], Threader1 und Threader2 [168] berechnet wurden. Die Wahl der Parameter der verschiedenen Methoden erfolgte, wie bereits in Abschnitt 8.3.3 beschrieben.
In der Abbildung 8.8 sind die Alignments nach ihrer Güte bezüglich des *RMS*–Kriteriums als Histogramm aufgetragen. Die dunkelblauen Balken zeigen die *RMS*–Abweichung für die aus JOY ausgewählten Referenzalignments. Gemäß eines der Kriterien für die Auswahl eines Alignments liegt die *RMS*–Abweichung aller 59 Referenzalignments zwischen 0 und 3Å, wobei die meisten Alignments eine *RMS*–Abweichung von etwa 2Å haben (siehe auch Tabelle 8.2).
Die mit 123D für diese Beispiele berechneten Alignments sind in den hellblau-

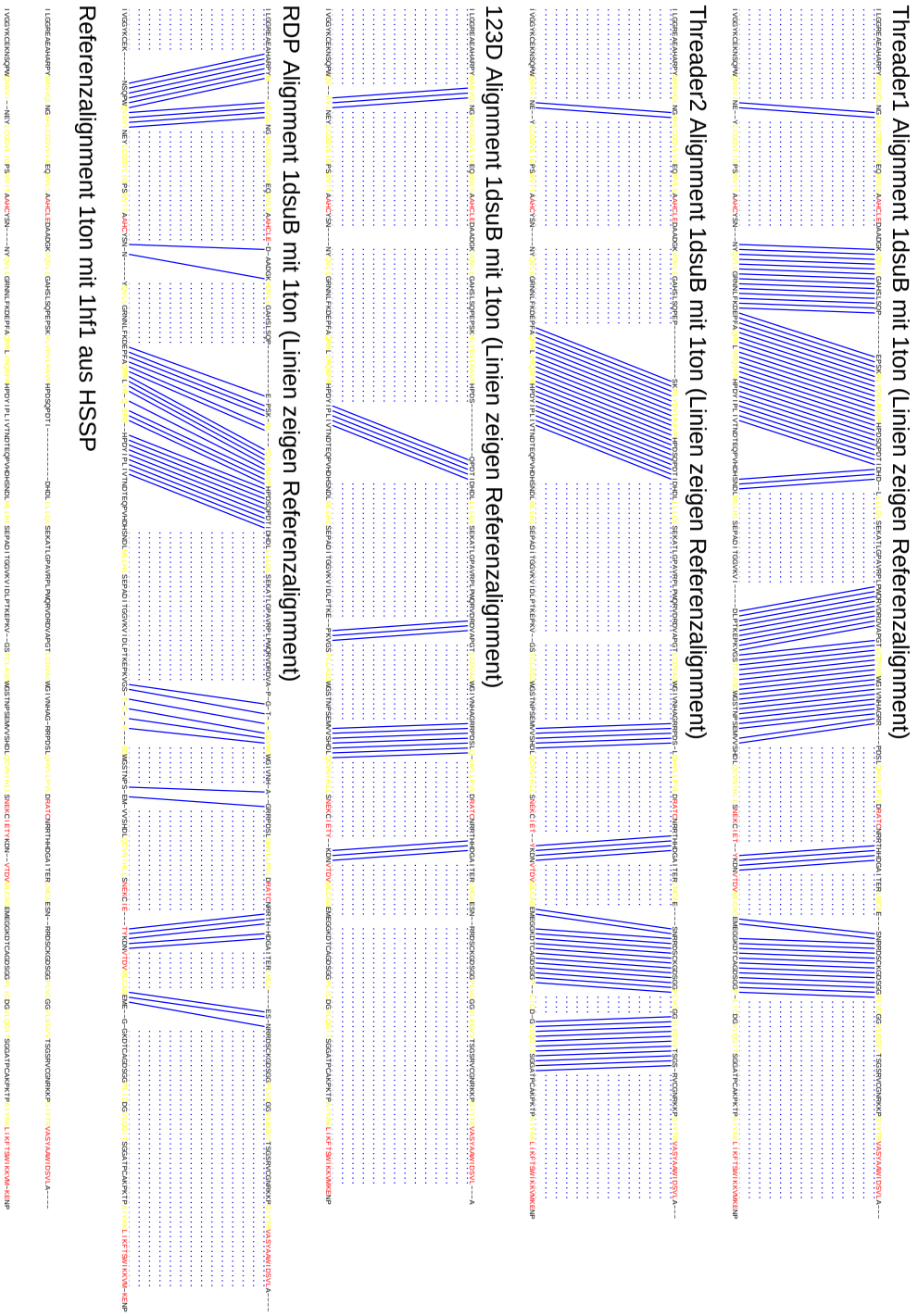


Abbildung 8.7: Positioneller Vergleich der Alignments der Serinproteasen `1dsuB` und `1ton`: `Threader1` (oben, *RMS* 6.8Å, 207), `Threader2` (*RMS* 5.9Å, 207), `123D` (*RMS* 2.63Å, 218) und RDP (*RMS* 4.9Å, 188). Die Linien zeigen die Verschiebung bezüglich des `HSSP`–Alignments (unten, *RMS* 2.7Å, 215).

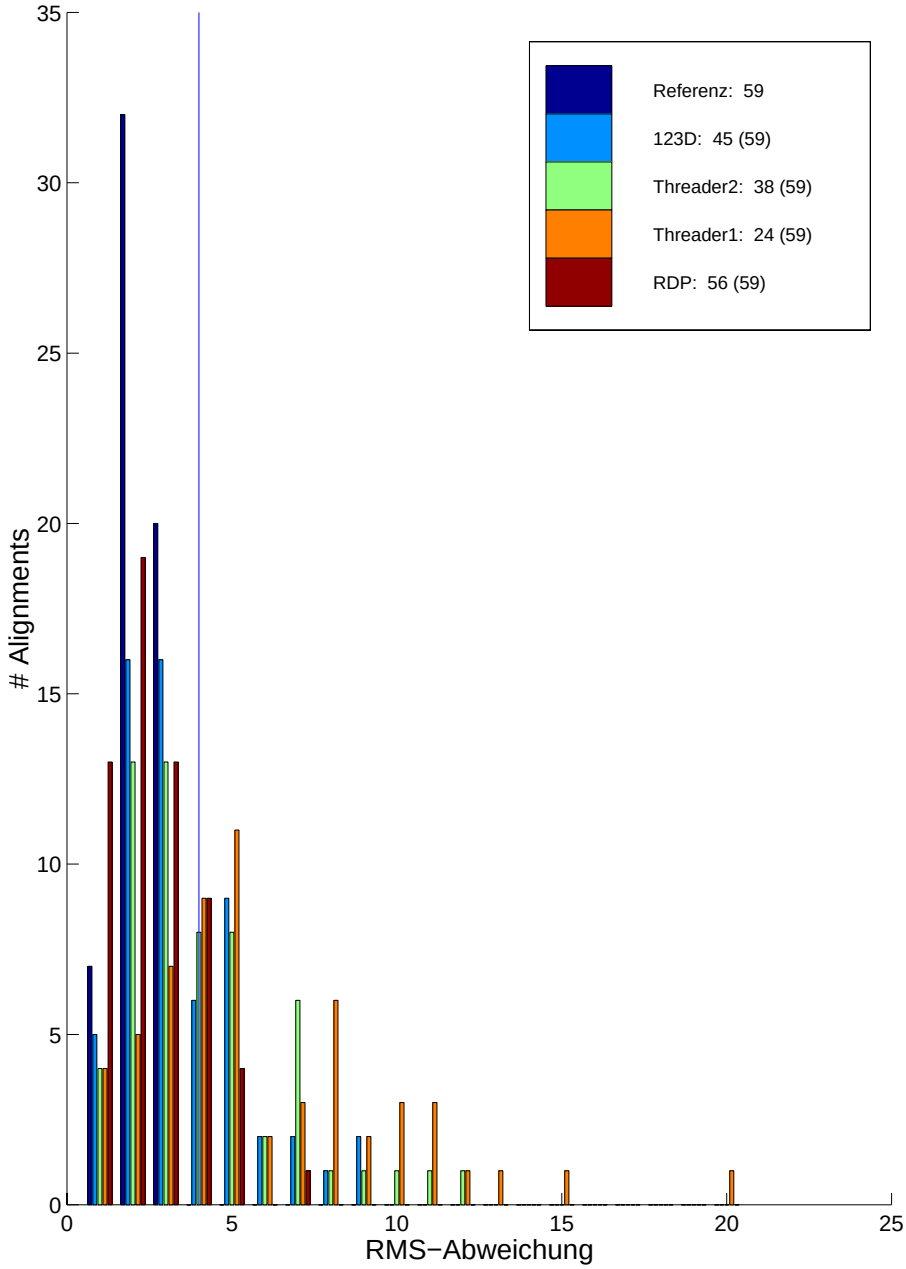


Abbildung 8.8: Histogramm der RMS–Werte der von 123D(hellblau), Threader2(grün), Threader1(orange), und RDP(rot) berechneten Sequenzstrukturalignments im Vergleich für 59 JOY (blau) Alignments. Die Legende gibt für jede Methode die Anzahl der Alignments (gesamt) an, deren *RMS*–Abweichung kleiner als die des JOY–Alignments + 2Å ist.

en Balken zusammengefaßt. In 45 Fällen überschreitet das 123D–Alignment die durch die Referenz vorgegebene *RMS*–Abweichung um weniger als 2Å. In 14 Fällen ist die *RMS*–Abweichung für das 123D–Alignment mehr als 2Å größer als

die des Referenzalignments. In 16 Fällen überschreitet die *RMS*–Abweichung für das `123D`–Alignment die zur Orientierung eingezeichnete 4Å–Marke.
Von den mit `Threader2` berechneten Alignments erreichen nur 38 der Alignments das gesetzte Qualitätskriterium und, für `Threader1` sind es sogar nur 24 von 59. Auch für die Beispiele aus der `JOY`–Datenbank erreicht die RDP–Methode mit den in Abschnitt 7.7 festgelegten Parametern die besten Ergebnisse. Und zwar liegt die *RMS*–Abweichung für 56 der 59 Alignments innerhalb der Toleranzgrenze von 2Å über der des Referenzalignments.
Abbildung 8.9 zeigt, daß nicht nur mehr RDP–Alignments das gesetzte Qualitätskriterium erfüllen, sondern daß auch in der Regel deren *RMS*–Abweichung geringer als die mit Konkurrenzmethoden berechneter Alignments ist. Es gibt vier Ausnahmen.
Die erste bilden die Nitrogenasen `1st3` (269) und `1sbc` (274), wo die RDP–Methode nur die 119 Reste beider Proteine mit einem *RMS* von 2.5Å aliniert, während alle anderen Methoden wie auch die Referenz zusätzlich auch die restlichen 129 Reste mit einer *RMS*–Abweichung von 1.0Å alinieren. Daß RDP nur den jeweils zweiten Teil der Proteine aliniert, liegt daran, daß in der RDP–Methode Gaps nicht bestraft werden. Auch beim Alignment einer weiteren Nitrogenase (`1sbt`, 275 Reste) gegen die Struktur von `1st3` ist RDP mit 3.5Å *RMS*–Abweichung schlechter als die anderen Methoden, da in dem Alignment eine Helix und damit auch der nachfolgende Strang um eine Helixwindung verschoben werden. Die Vermutung, daß die Struktur `1st3` eine Besonderheit aufweist, die diesen Fehler hervorruft, konnte nicht belegt werden.
Die Protease `1bbs` fällt ebenfalls zweifach auf: Zum einen ist das `123D`–Alignment (3.6Å) von `1bbs` mit der Sequenz von `2apr` von der *RMS*–Abweichung 0.5Å besser als das entsprechende RDP–Alignment (4.1Å), das im wesentlichen im hinteren Bereich einige kleinere Verschiebungsfehler in den Strängen aufweist. Zum anderen gehört das RDP–Alignment der Säureproteasen `1mpp` (357) und `1bbs` (331) zu den drei RDP–Alignments, die das Gütekriterium nicht erreichen. Das RDP–Alignment hat bei einer Sequenzidentität von 23.5% eine *RMS*–Abweichung von 4.7Å (in der Referenz: 3.0Å bei 30% Sequenzidentität).
Das zweite „schlechte“ RDP–Alignment aliniert die zwei Hydrolasen `1ezm` (298) und `8tlnE` (316) mit einer *RMS*–Abweichung von 4.8Å (Referenzalignment: 2.9Å und 31.0%). Hier erreichen auch die anderen Methoden kein besseres Ergebnis (siehe zweitletzte Spalte in Abbildung 8.9).
Das schlechteste Alignment von RDP wird für die zwei Flavodoxine `3fxn` (138 Reste) und `2fcr` (173) mit einer *RMS*–Abweichung von 4.3Å berechnet. Obwohl das Strukturalignment eine *RMS*–Abweichung von 2.6Å hat, liefern hier anscheinend alle Sequenzstrukturalignmentmethoden die gleichen oder zumindest gleich schlechte Alignments (siehe letzte Spalte in Abbildung 8.9). Es scheint sich hierbei um ein für Sequenzstrukturalignmentmethoden besonders schwieriges Beispiel zu handeln. Ein Grund ist dafür die Tatsache, daß zum strukturrichtigen Alignment vor dem in beiden Proteinen letzten Strang–Helix–Abschnitt Proteinen bezüglich

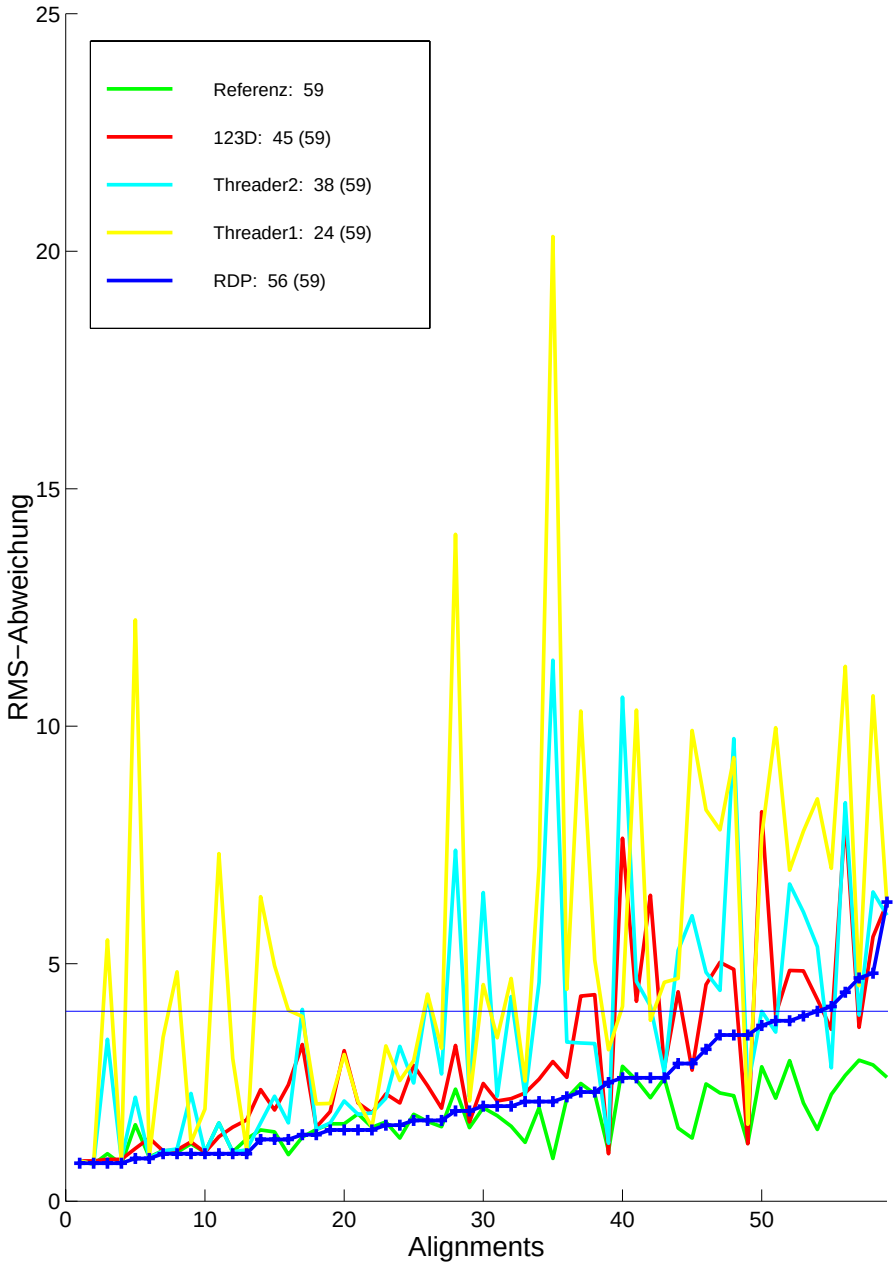


Abbildung 8.9: Vergleich der *RMS*–Abweichungen der von `123D` (rot), `Threader2` (hellblau), `Threader1` (gelb) und RDP (blau) berechneten Sequenzstrukturalignments für 59 `JOY` (grün) Strukturalignments (aufsteigend sortiert nach RDP).

der Struktur `3fxn` eine 21 Reste lange Insertion eingefügt werden muß.

Insgesamt liefert jedoch die RDP–Methode auch für die aus der `JOY`–Datenbank extrahierte Testmenge deutlich bessere Sequenzstrukturalignments als die zum Vergleich herangezogenen Methoden `Threader` und `123D`.

8.3.5 Vergleich auf der Basis von SARF–Strukturalignments

Die Abbildungen 8.10 und 8.11 vergleichen die Güte der mit der RDP–Methode für die mittels SARF2 ausgewählten Sequenz–Strukturpaarungen berechneten Alignments mit denen, die mit den im Abschnitt 8.3.3 beschriebenen Methoden und Parametrisierungen berechnet wurden. Die RDP–Methode berechnet für 45 der 73 wie oben beschrieben mit SARF2 bestimmten Paare aus Struktur und Sequenz ein Sequenzstrukturalignment, dessen *RMS*–Abweichung innerhalb der vorgegebenen Toleranzgrenze von 2Å über der *RMS*–Abweichung des zugehörigen Strukturalignments liegt. Dies sind fast doppelt so viele „gute“ Alignments, wie 123D mit 24 Alignments innerhalb der Toleranzgrenze für den besten Parametersatz aus 30 sinnvollen Parameterkombinationen berechnet. Wie auch in den vorangegangenen Tests belegen die Methoden Threader2 und Threader1 mit 20 beziehungsweise 17 „guten“ Sequenzstrukturalignments abgeschlagen die Plätze drei und vier (siehe Legenden von Abbildung 8.10).

In der Abbildung 8.10 sind die Alignments nach ihrer Güte bezüglich des *RMS*–Kriteriums als Histogramm aufgetragen. Die dunkelblauen Balken zeigen die *RMS*–Abweichung für die mittels SARF2 berechneten Strukturalignments, die im folgenden als Referenz dienen. Gemäß eines der Kriterien für die Auswahl eines Alignments liegt die *RMS*–Abweichung aller 73 Referenzalignments zwischen 1.3 und 2.8Å, wobei die meisten Alignments eine *RMS*–Abweichung von etwa 2Å haben (siehe auch Tabelle 8.3).

Threader1 ist auch auf dieser Testmenge deutlich schlechter als Threader2. So haben 39 der Threader1–Alignments eine *RMS*–Abweichung über 6Å, 27 über 10Å und 2 sogar über 20Å. Für Threader2 überschreiten zwar immer noch 36 Alignments die 6Å–Marke, doch nur noch 22(0) Alignments haben eine *RMS*–Abweichung größer als 10(20)Å. Die Verteilungen der *RMS*–Abweichungen der von Threader1 und Threader2 berechneten Alignments sind in Abbildung 8.10 in den orangen beziehungsweise grünen Balken gezeigt.

Alignments mit einer *RMS*–Abweichung größer als 10Å sind in aller Regel so falsch, daß sie keinesfalls als Basis für eine anschließende vergleichende Modellierung verwendet werden können. Zum Vergleich fallen in diese Kategorie der vollkommen falschen Alignments nur 12 der 123D–Alignments und sogar nur 9 der RDP–Alignments. Der qualitative Unterschied zwischen den an der GMD entwickelten Methoden und Threader wird auf dieser Testmenge also bereits an der Anzahl der vollkommen falschen Sequenzstrukturalignments sehr deutlich. Dagegen unterscheiden sich die 123D–Methode und die RDP–Methode, wie Abbildung 8.10 zeigt, eher in der Güte der eventuell brauchbaren Alignments. So haben nur 23 von 73 RDP–Alignments eine *RMS*–Abweichung größer als 5Å, während für 30 123D–Alignments diese Marke überschritten wird. Folgerichtig liegt auch die mittlere *RMS*–Abweichung der RDP–Alignments bei 4.8 und die der 123D–Alignments bei 6.5 Å. Für Threader1 und Threader2 beträgt der Mittelwert der *RMS*–Abweichungen sogar 8.7Å beziehungsweise 7.6Å.

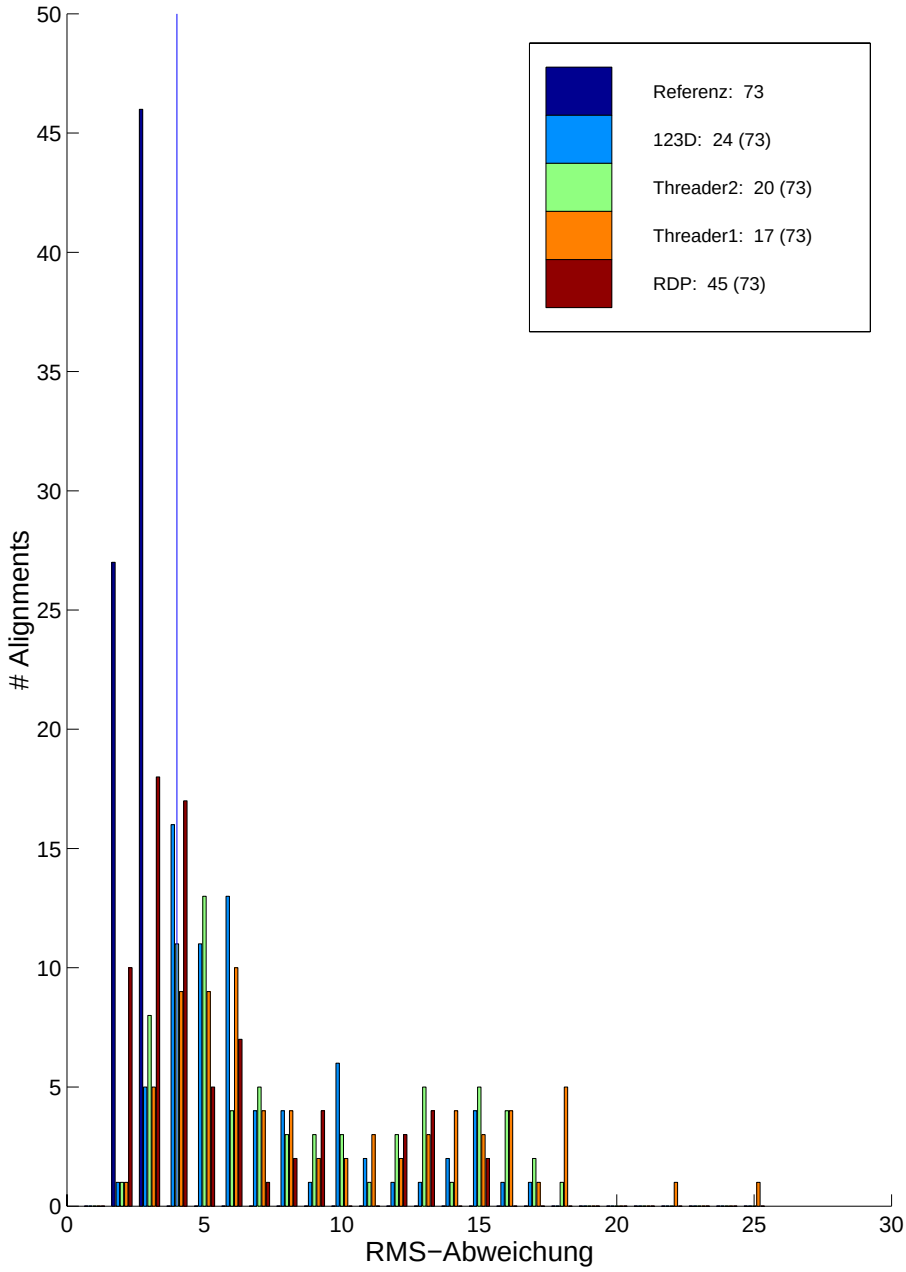


Abbildung 8.10: Histogramm der *RMS*–Abweichungen der von 123D (hellblau), `Threader2` (grün), `Threader1` (orange) und RDP (rot) berechneten Sequenzstrukturalignments im Vergleich für 73 `SARF2` (dunkelblau) Strukturalignments. Die Legende gibt für jede Methode die Anzahl der Alignments (Gesamtanzahl) an, deren *RMS*–Abweichung kleiner als die des `SARF2`–Strukturalignments + 2Å ist.

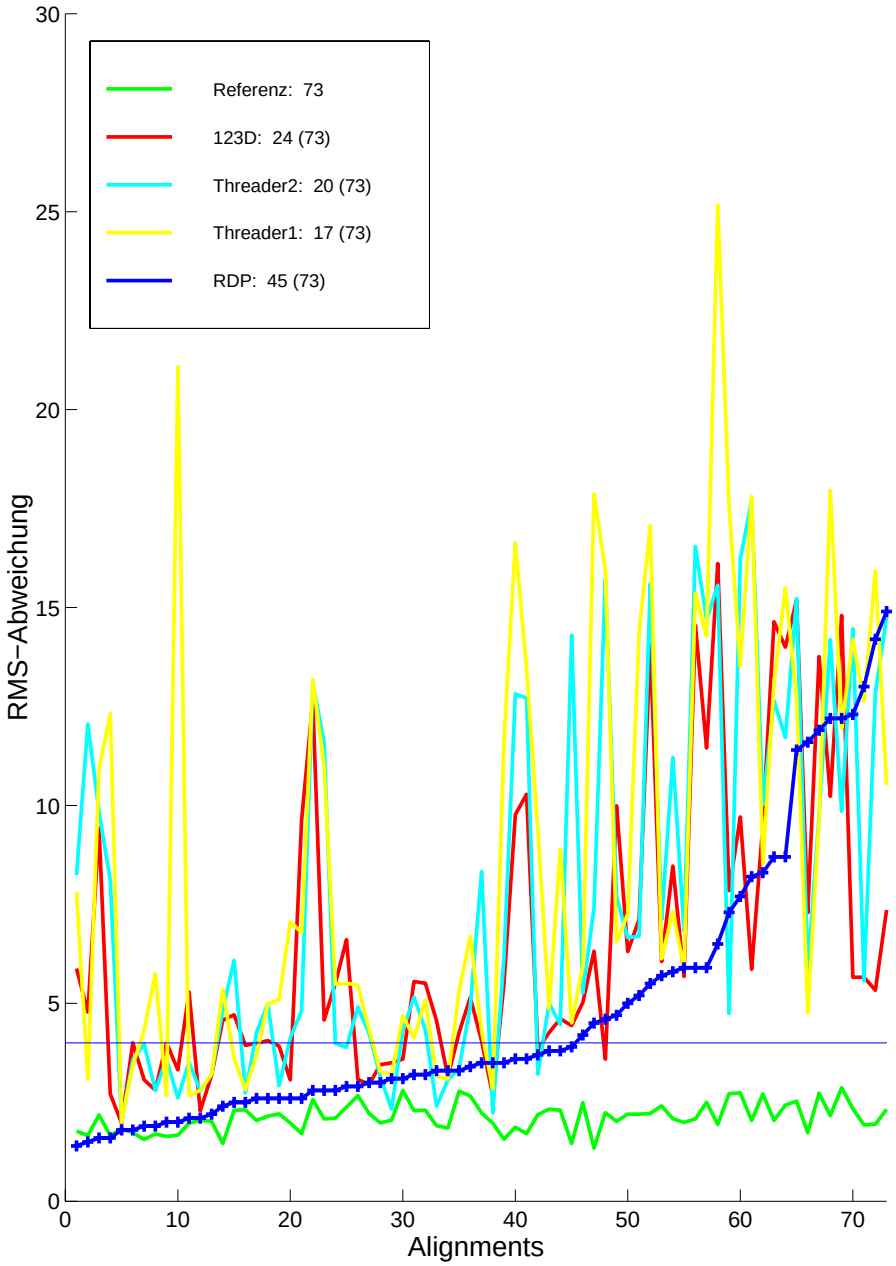


Abbildung 8.11: Vergleich der *RMS*–Abweichungen der von `123D` (rot), `Threader2` (hellblau), `Threader1` (gelb) und RDP (blau) berechneten Sequenzstrukturalignments für 73 `SARF2` (grün) Strukturalignments (aufsteigend sortiert nach RDP).

In Abbildung 8.11 sind *RMS*–Abweichungen der einzelnen Alignments aufsteigend sortiert nach der *RMS*–Abweichung des jeweiligen RDP–Alignments aufgetragen. Die Abbildung zeigt sehr deutlich, daß das RDP–Alignment in den meisten Fällen besser ist als das mit den Vergleichsmethoden berechnete Alignment. In Abbildung 8.11 findet sich erst an Position 59 die erste ernsthafte Abweichung von die-

sem Trend. RDP aliniert die Sequenz von 256bA gegen die Struktur von 1was mit einer *RMS*–Abweichung von 7.3Å bei 90 zugeordneten Positionen. Nur Threader2 berechnet für die beiden Proteine mit einer Vier–Helix–Bündel–Faltung mit 4.7Å ein besseres Alignment. In der ersten und dritten Helix weisen beide Alignments die gleichen Fehler im Vergleich mit der Strukturuperposition auf, nur in der zweiten Helix gelingt es Threader2 im Unterschied zu RDP 11 zusätzliche Reste strukturrichtig zuzuordnen, so daß sich die bessere *RMS*–Abweichung ergibt.
Im Fall der Lipocaline 1hbq und 1bbpA (Position 61 in Abbildung 8.11) ist nur das 123D–Alignment besser als RDP (RDP–Alignment 8.2Å und 123D 5.9Å). Dies gilt jedoch nur für den Fall, daß 1bbpA als Struktur und 1hbq als Sequenz angenommen werden. Vertauscht man für diese beiden Lipocaline Struktur und Sequenz, so berechnet auch die RDP–Methode ein Sequenzstrukturalignment mit einer *RMS*–Abweichung von 5.9Å. Abbildung 8.12 vergleicht die zwei Alignments, indem die Alignments bezüglich der Sequenz von 1bbpA (jeweils in der dritten Zeile) zusammen mit den zugehörigen Sekundärstrukturen ausgerichtet werden. Der wesentliche Unterschied zwischen den zwei Alignments besteht darin, daß bei

```
Struktur 1hbq_000:E...---RDCRVSSFRVK-ENFDKARFAGTWYAMAK---KDPEG--------LFLQDN-IVAEFSVD---
SECSTR   1hbq_000:           hhh              eeeeeeee   e                ee eeeeeee
Seq/Str 1bbpA_000:....NVYHDGACPEVKPV-DNFDWSNYHGKWWEVAKYPN----S-----VEK--YGKC-GWAEYTPEGKS
SECSTR  1bbpA_000:     eeee              hhh  eeeeeeee                    ee eeeeeee
Sequenz  1hbq_000:.ERD.....CRVSSFRV-KENFDKARFAGTWYAMAK--------KDPEGLFL--QD--NIVAEFSVD---
SECSTR   1hbq_000:           hhh              eeeeeeee        e          e  eeeeeeee

Struktur 1hbq_070:ENGQMS-----------ATAKGRVRLLNN--WDVCADMVGTFTDTEDPAKFKMKYWG----VASFLQKGN
SECSTR   1hbq_070:    ee           eeeeeeee       eeeeeeeeeeee     eeeeeeee           ee
Seq/Str 1bbpA_070:-----V-----------KVSNYHVIHGKE--YFIEGTAYPVGDSKIGKIYHKLTYGGVTKE---------
SECSTR  1bbpA_070:     e           eeeeeeee  ee  eeeeeeeee       eeeeeeee  eeee
Sequenz  1hbq_070:------ENGQMSATAKGRVRLL-------NNWDVCADMVGTFTDTEDPAKFKMKYWGVAS----------
SECSTR   1hbq_070:          eeeeeeeeee            eeeeeeeeeee     eeeeeeee

Struktur 1hbq_140:DD--------H-WIIDTDYETFAVQYSC-RLLNLDGTCADSYSFVFARD---PSGFSPE----VQKIVRQ
SECSTR   1hbq_140:ee        e eeeee    eeeeeee eee     eeeeeeeeee          hh    hhhhhhh
Seq/Str 1bbpA_140:NV--------FN-VLSTDNKNYIIGYYCK--YDEDKKGHQDFVWVLSRSKVL----TGE----AKTAVEN
SECSTR  1bbpA_140:ee        ee eeee    eeeeeeee  ee    eeeeeeeeee          hh    hhhhhhh
Sequenz  1hbq_140:--FLQKGNDDHW-IIDTDYETFAVQYSCR--LLNLDGTCADSYSFVFARDPS----G--FSPEVQKIVRQ
SECSTR   1hbq_140:      eeeeee eeee    eeeeeeee  ee     eeeeeeeeee             hhhhhhhhh

Struktur 1hbq_210:RQEELC---------LARQYRLI............PHNGYCNGKSERNIL
SECSTR   1hbq_210:hhhh
Seq/Str 1bbpA_210:YLIGSP---------VVDSQKLVYSDFSEAACKVN.........-----.
SECSTR  1bbpA_210:hhhh              hhh       hhhh
Sequenz  1hbq_210:RQE---ELCLARQYRLIPHNGYC............NGKSERNIL-----.
SECSTR   1hbq_210:hhh   h
```

Abbildung 8.12: Vergleich der RDP–Sequenzstrukturalignments von Struktur 1hbq (1. Zeilen) mit Sequenz 1bbpA (3. Zeilen) und von Struktur 1bbpA (3. Zeilen) mit Sequenz 1hbq (5. Zeilen).

Verwendung von 1hbq als Struktur der dritte Strang korrekt gegen den dritten Strang von 1bbpA aliniert wird (*RMS*–Abweichung 5.9Å), während dieser bei Verwendung von 1bbpA als Struktur um 5 Reste verschoben wird, was eine *RMS*–Abweichung von 8.2Å zur Folge hat.
Da beide Lipocaline in etwa die gleiche Länge haben (1bbpA 173, 1hbq 183 Reste), muß dieser Unterschied in den Alignments auf Unterschiede in strukturellen Eigenschaften der beiden Strukturen zurückzuführen sein. Beide Strukturen sind mit einer Auflösung von kleiner gleich 2.0Å aufgelöst und auch ihre native Paarpo-

tentialbewertung für das verwendete auf Voronoikontakten basierende Potential `ED6SD6` liegt mit −62.8 für `1bbpA` und −64.0 für `1hbq` in der gleichen Größenordnung, so daß Fehler in den Strukturen, die diesen Alignmentfehler hervorrufen könnten, nahezu auszuschließen sind. Also wird der Fehler wahrscheinlich entweder durch Unterschiede im Kontaktnetzwerk oder dadurch hervorgerufen, daß der betroffene Sequenzabschnitt von `1hbq` bezüglich des Paarpotentials nicht so gut in den korrespondierenden Strukturabschnitt von `1bbpA` paßt, wie dies umgekehrt der Fall ist.
Für die weiteren RDP–Sequenzstrukturalignments, die eine *RMS*–Abweichung zwischen 5 und 9Å haben (Positionen 51 bis 64 in Abbildung 8.11) und damit bedingt verwendbar sind, aber falsche Zuordnungen enthalten, ist das RDP–Alignment besser als die von den Konkurrenzprogrammen berechneten Alignments.
Für 5 der verbleibenden 9 Paare, die von der RDP–Methode *RMS*–Abweichungen größer als 10Å aliniert werden, ist dies nicht der Fall, wie Abbildung 8.11 zeigt. Von keiner der zum Vergleich eingesetzten Methoden wird jedoch ein wirklich gutes Alignment für diese Beispiele berechnet. In zwei Fällen wird von `Threader2` und in vier Fällen von `123D` ein bedingt verwendbares Alignment mit einer *RMS*–Abweichung unter 7.5Å berechnet.

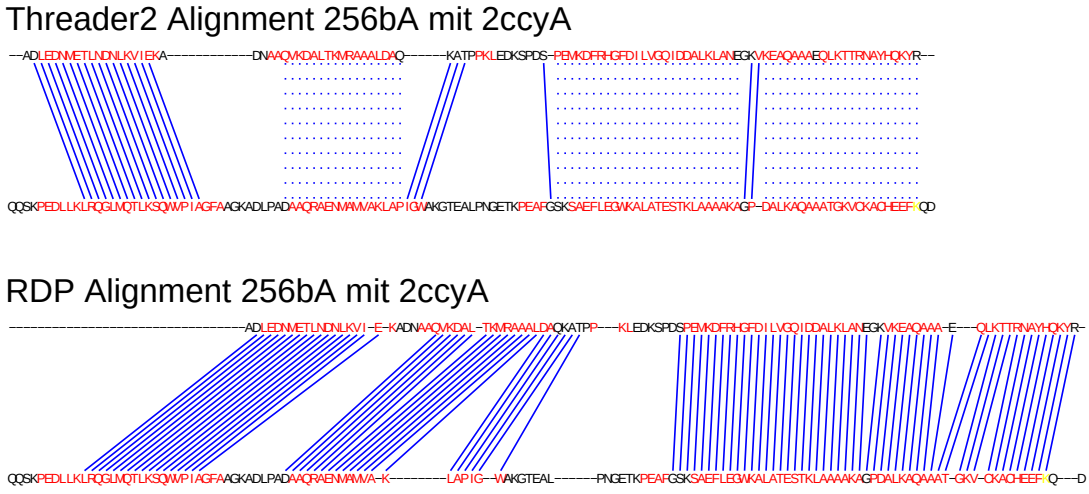


Abbildung 8.13: Vergleich der mit `Threader2` und RDP berechneten Alignments von Struktur `256bA` mit Sequenz `2ccyA` gegen das mit `SARF2` berechnete Strukturalignment

Bei dem ersten Beispiel, für das `Threader2` ein besseres Alignment als RDP

liefert, handelt es sich um die Proteine mit Vier–Helix–Bündel–Faltung `256bA` und `2ccyA`. Abbildung 8.13 vergleicht `Threader2`–Alignment (5.8Å) und RDP–Alignment (11.6Å) gegen das mit `SARF2` berechnete Strukturalignment. Von dem `Threader2`–Alignment werden die zweite bis vierte Helix korrekt zugeordnet und nur die erste Helix wird um einige Positionen verschoben. Dagegen ordnet RDP die erste Helix von `256bA` der zweiten von `2ccyA` zu und hat auch in den korrekt zugeordneten dritten und vierten Helices kleinere Verschiebungsfehler, der eventuell aber bereits aus dem ersten Fehler resultiert.
Bei den Proteinen `1prtD` und `1prtF` mit einer Faltung, die als *OB–Fold* bezeichnet wird, alinieren `Threader2` und `123D` drei der fünf Stränge korrekt, während dies RDP für einen der Stränge ohne Verschiebungsfehler gelingt, was den Unterschied in der *RMS*–Abweichung (`Threader2` und `123D` 5.6Å, RDP 13.0Å) erklärt.
Die Strukturen des Hisactophilins `1hce` und des Interleukins–1 `1i1b` bilden das zweite Beispiel, wo das `123D`–Alignment (5.7Å *RMS*) signifikant besser ist als das RDP–Alignment (12.3Å *RMS*, Position 70 in Abbildung 8.11). In diesem Beispiel führt auch ein Vertauschen von Sequenz und Struktur nicht zu einem besseren RDP–Alignment. Das RDP–Alignment der Sequenz `1i1b` gegen die NMR–Struktur `1hce` hat zwar eine sehr niedrige Paarpotentialbewertung (-30.3 gegenüber -30.0 für das native Protein `1hce`), dennoch ist das Alignment ebenso falsch wie das für die Ausgangspaarung. Das korrekte und in diesem Fall sehr kompakte Strukturalignment erzielt dagegen nur eine Paarpotentialbewertung von -26.0. Dies legt die Vermutung nahe, daß bei diesem Beispiel die RDP–Methode von der verwendeten Bewertungsfunktion fehl geleitet wird und somit keine Möglichkeit hat, die korrekte Zuordnung zu finden.
An Position 72 in Abbildung 8.11 befindet sich das RDP–Alignment der zwei Lipocaline `1hbq` und `1epaB` mit `1hbq` als Struktur mit einer *RMS*–Abweichung von 14.2Å. Andererseits gelingt es der RDP–Methode die Sequenz `1hbq` so auf die Struktur `1epaB` abzubilden, daß die zugehörige Superposition eine *RMS*–Abweichung von 5.9Å hat. Dies verstärkt die bereits oben geäußerte Vermutung, daß die Struktur `1hbq` eine Besonderheit bezüglich des Paarpotentials aufweist. Erstaunlicherweise tritt der gleiche Effekt für `123D` genau in der umgekehrten Richtung auf. Verwendet man für `123D` `1hbq` als Struktur, erhält man ein Alignment mit 5.3Å *RMS*–Abweichung, verwendet man hingegen `1epaB` als Struktur, so beträgt die *RMS*–Abweichung des `123D`–Alignments 14.6Å.
Das Interleukin–1 `1i1b` ist als Struktur auch an dem vierten Beispiel beteiligt, wo das `123D`–Alignment (7.4Å) mit der Sequenz des Serinproteaseinhibitors `1wbc` (7.4Å) besser ist als das zugehörige RDP–Alignment (14.9Å). Beide Proteine haben eine Faltung vom Typ *β–Trefoil* und `SARF2` superpositioniert 115 Reste mit einer *RMS*–Abweichung von nur 2.3Å (siehe Tabelle 8.3. Das RDP–Alignment dieser beiden Strukturen ist gleichzeitig auch das RDP–Alignment mit der höchsten *RMS*–Abweichung und befindet sich daher an Position 73 in Abbildung 8.11). Wie bereits im ersten Alignment, an dem die Interleukinstruktur beteiligt war, schafft es die RDP–Methode auch hier nicht, die Faltblätter der beiden Strukturen

korrekt einander zuzuordnen. In dem RDP–Alignment werden zwar β–Stränge β–Strängen zugeordnet, doch wird bereits dem ersten von zwölf Strängen aus `1i1b` der vierte aus `1wbc` zugeordnet. Diese große Verschiebung um 57 Reste hat die hohe *RMS*–Abweichung zur Folge. Im `123D`–Alignment wird eine derartige Verschiebung dagegen bereits durch die Kosten für die aus der Verschiebung resultierenden Gaps verhindert, während in der RDP–Methode Insertionen und Deletionen kaum bestraft werden.
Die Diskussion der wenigen schlechten RDP–Alignments zeigt, daß auch die RDP–Methode nicht in allen Fällen für entfernt verwandte oder strukturell ähnliche Proteine direkt verwendbare Sequenzstrukturalignments berechnen kann. In einigen wenigen Fälle liefern sogar andere Methoden bessere Ergebnisse. Die Gründe dafür sind unterschiedlich. In einigen Beispielen scheint der Grund in dem durch die spezielle Struktur definierten Kontaktnetzwerk zu liegen, in anderen in der Tatsache, daß die RDP–Methode so gut wie keine Gapkosten verwendet. Eine weitere Fehlerursache besteht darin, daß die Kostenfunktion, insbesondere der Paarpotentialanteil, unzureichend genau ist. Was dazu führt, daß die von der RDP–Methode berechneten Sequenzstrukturalignments von der Kostenfunktion weitaus besser bewertet werden, als dies für ein zugehöriges strukturell richtiges Alignment der Fall ist.
Eine Optimierungsmethode für Sequenzstrukturalignments kann aber immer nur so gut sein, wie es durch die verwendete Kostenfunktion vorgegeben wird. Aus diesem Grund soll insbesondere der Paarpotentialanteil der Kostenfunktion im Anschluß an diese Arbeit, unter anderem anhand der hier identifizierten Problembeispiele, genauer untersucht werden. Die zahlreichen Vorschläge für die Ableitung von Paarpotentialen in der Literatur (siehe Abschnitt 5.2.4) belegen, daß für dieses Problem bislang keine Lösung gefunden wurde, deren Verwendung garantiert, daß die optimale Lösung des Alignmentproblems identisch oder zumindest sehr ähnlich der biologisch beziehungsweise strukturell richtigen Lösung ist. Eine genauere Analyse der durch die Unzulänglichkeiten der Kostenfunktion hervorgerufenen Fehler würde den Rahmen dieser Arbeit sprengen.
Insgesamt gesehen liefert – wie Abbildung 8.11 deutlich zeigt – die RDP–Methode für wesentlich mehr Beispiele der Menge der mit `SARF2` aus der Repräsentativmenge `hobohm96_25` abgeleiteten, damit nur strukturell ähnlichen und nicht sequenzhomologen Paare *gute* (*RMS*–Abweichung ≤ 5Å) und *brauchbare* (*RMS*–Abweichung ≤ 10Å) Sequenzstrukturalignments.
Damit konnte nachgewiesen werden, daß mit der RDP–Methode eine wesentliche Verbesserung der Alignmentqualität sowohl für Sequenzstrukturpaare mit erkennbarer Sequenzähnlichkeit als auch für Paare, wo eine evolutionäre Verwandschaft nicht auf Sequenzebene erkennbar ist, erzielt werden kann. Die mit RDP erzielten Verbesserungen tragen somit nicht nur dazu bei, daß die Verfahren der vergleichenden Modellierung auf sehr entfernt verwandte Proteine ausgedehnt werden können, sondern sie steigern auch dort die Modellqualität, wo diese Verfahren schon heute standardmäßig eingesetzt werden.

8.3.6 Vergleich der Laufzeiten

Abbildung 8.3.5 vergleicht die Laufzeiten der bei der Analyse der Alignmentqualität verwendeten Sequenzstrukturalignmentmethoden `123D`, `Threader1` und `Threader2` mit der Laufzeit der in dieser Arbeit entwickelten RDP–Methode auf der Menge der mittels `SARF` bestimmten Sequenzstrukturpaare (siehe Abschnitt 8.3.2). Diese Testmenge wurde für den Laufzeitvergleich ausgewählt, da

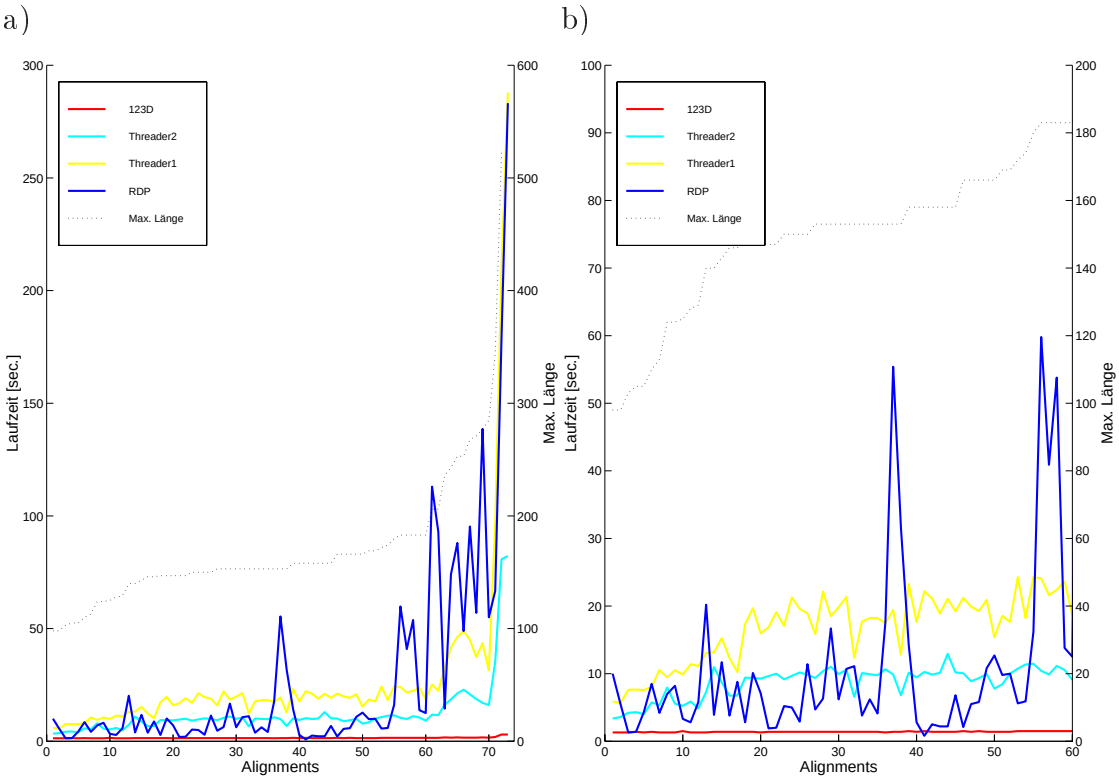


Abbildung 8.14: Vergleich der Laufzeiten anhand der 73 der `SARF`–Testmenge aus Abschnitt 8.3.2: Die Laufzeiten sind sortiert nach der maximalen Länge von Struktur und Sequenz (gepunktete Linie rechte Y-ordinate). Teil b) zeigt einen vergrößerten Auschnitt für die 60 Sequenzstrukturpaare mit max. Länge kleiner 200.

sie den typischen Anwendungsfall von Sequenzstrukturalignmentmethoden repräsentiert. Da zwischen den Sequenzstrukturpaaren dieser Testmenge keine signifikanten Sequenzähnlichkeit vorhanden sind, ist außerdem zu erwarten, daß die lokalen Ähnlichkeiten, nach denen die RDP–Methode sucht, nur schwer aufzudecken sind und damit die Laufzeit einer Methode wie RDP entsprechend größer als die der anderen, nicht rekursiv arbeitenden Methoden ist.
Da die `123D`–Methode auf der einfachen dynamischen Programmierung beruht, ist `123D` erwartungsgemäß die schnellste Methode und gegen die Schwierigkeit des jeweiligen Beispiels am unempfindlichsten. Die Tatsache, daß die Laufzeit

von 123D asymptotisch mit dem Produkt der Längen von Sequenz und Struktur wächst, ist in der vorgegebenen Skalierung nur im hinteren Bereich vom Teil a) der Abbildung 8.14 zu erkennen. Ansonsten überwiegt bei der schnellen Methode der Zeitanteil, der für das Laden des Programms und der Daten benötigt wird.
Die für Threader zu erwartende Laufzeit ist aufgrund der doppelten dynamischen Programmierung quadratisch in dem Produkt der Längen der Eingabe. Im Unterschied zu 123D nimmt daher die Laufzeit wesentlich schneller mit der Länge der Eingabe zu. Anscheinend spielen aber auch bei der Threader–Laufzeit andere Faktoren neben der Eingabelänge eine Rolle, da die gelbe und hellblaue Kurve dadurch nicht ganz erklärt werden.
Insbesondere für die schwierigeren Fälle sind sowohl beide Threader–Versionen als auch RDP langsamer als die sehr schnelle 123D–Methode. Für überraschend viele Fälle (etwa die Hälfte der Beispiele) ist die RDP–Methode sogar schneller als der auf dynamischer Programmierung basierende Threader. Insgesamt ist der Zeitbedarf der RDP–Methode in der gleichen Größenordnung wie der von Threader. Der große Unterschied besteht jedoch darin, daß die RDP–Methode, wie im vorangegangen Abschnitt gezeigt wurde, in der gleichen Zeit wesentlich bessere Sequenzstrukturalignments berechnet als Threader.
Abbildung 8.14 zeigt aber auch, daß die Laufzeit der RDP–Methode stark mit der der maximalen Länge der zu alinierenden Proteine zu nimmt. Dieser Effekt wirkt sich nicht so sehr aus, wenn die Zielsetzung die Berechnung genauer Sequenzstrukturalignments zwischen eindomänigen Proteinen oder Proteinendomänen ist, deren Länge in der Regel kleiner als 300 ist. Der Effekt kommt vielmehr zum Tragen, wenn Datenbanksuchen gegen repräsentative Proteinstrukturmengen durchgeführt werden, die typischerweise aus 600 bis 1000 Proteinstrukturen unterschiedlichster Länge (bis über 1000 Aminosäuren) bestehen.
Eine Datenbanksuche mit einer typischen Sequenz der Länge 158 aus dem gegenwärtigen CASP 3–Wettbewerb gegen eine 632 elementige Repräsentativmenge benötigt so zum Beispiel ungefähr fünf Stunden, während der gleiche Suchlauf von 123D in einigen Minuten durchgeführt werden kann. Dabei ist zu berücksichtigen, daß die RDP–Methode bisher hauptsächlich hinsichtlich Alignmentqualität und weniger hinsichtlich der Laufzeit optimiert wurde.
Erste Analysen haben gezeigt, daß die RDP–Methode dabei die meiste Zeit für die Berechnung von Sequenzstrukturalignments zwischen Proteinen verbraucht, für die sich später heraus stellt, daß sie nicht ähnlich sind. Ein Verbesserung, was dieses Problem betrifft, ist sicher durch die Entwicklung restriktiverer Signifikanz– und Zulässigkeitskriterien zu erreichen, die es erlauben, frühzeitiger Teilbäume des Lösungsbaumes abzuschneiden.
Wie bereits zu Beginn dieses Kapitels beschrieben, besteht ein sinnvolle Vorhersagestrategie ohnehin darin, zunächst mit der schnellen 123D–Methode die Faltungsklasse zu identifizieren, um dann mit der RDP–Methode auf der so eingeschränkten Suchmenge die beste Modellstruktur für die untersuchte Sequenz zu identifizieren und ein strukturgenaues Alignment zu berechnen. Dafür reicht die

Geschwindigkeit von RDP auch jetzt schon völlig aus, so däs RDP den Bereich der erfolgreich durchführbaren Proteinstrukturvorhersagen erweitert.

8.4 Faltungserkennung

8.4.1 Testmenge für die Erkennungsexperimente

Beispiele für echte Faltungserkennungstests sind schwer zu finden, da sie zwei Kriterien zu erfüllen haben:

- Zum einen sollte die Sequenzidentität zwischen den zu vergleichenden Proteinen niedrig (etwa kleiner 30%) sein.
- Zum anderen müssen die Strukturen dennoch so ähnlich sein, daß ihre Ähnlichkeiten von Sequenzstrukturalignment– beziehungsweise Faltungserkennungsmethoden erkannt werden können.

Das erste Kriterium ist einfach zu erfüllen, indem von einer Repräsentativmenge von Proteinen ausgegangen wird, deren Elemente im paarweisen Vergleich eine Sequenzidentität unterhalb der verwendeten Ähnlichkeitsschranke aufweisen. Weitaus schwieriger gestaltet sich die Gewährleistung des zweiten Kriteriums. Eine Erkennung von Ähnlichkeiten, die nicht auf Sequenzebene sondern nur auf Strukturebene zu Tage treten, kann man von Sequenzstrukturalignmentmethoden sicher nur dann erwarten, wenn ein genügend großer Teil beider Proteinstrukturen mit hinreichend kleiner *RMS*–Abweichung superpositionierbar ist.
Für die folgenden Untersuchungen wurden aus einer repräsentativen Proteinmenge (`hobohm96_25` [148], 486 Proteine) alle Proteine ausgewählt (251), die aus nur einer Domäne bestehen. Durch die Beschränkung auf eindomänige Proteine wurden spezielle Randeffekte ausgeschlossen, wie sie zum Beispiel durch die hydrophobe Packung an Domänschnittstellen auftreten. Im nächsten Schritt wurden die Proteine der so eingeschränkten Menge entsprechend der `SCOP`–Faltungsklassifikation [243] (siehe auch Abschnitt 3.3.5) annotiert und die Proteine mit gleicher Faltung zu sogenannten Faltungsfamilien zusammengefaßt.
Das Ergebnis der `SCOP`–Annotation sind elf Faltungsfamilien mit zwischen fünf und elf Mitgliedern pro Familie. Tabelle 8.5 zeigt die elf ausgewählten Faltungsfamilien, deren jeweilige Größe und die minimale und maximale Anzahl Reste der Familienmitglieder.
Die Tatsache, daß zwei Proteine von `SCOP` der gleichen Faltungsklasse zugeordnet werden, ist jedoch nicht in jedem Falle gleichbedeutend damit, daß alle Proteine der selben Faltungsklasse gut superpositionierbar sind. Abbildung 8.15 zeigt – getrennt für die elf Faltungsfamilien – für jedes mögliche Paar von Strukturen einer Faltungsfamilie die Anzahl der mit `SARF2` mit einer maximalen *RMS*–Abweichung von 3.4Å superpositionierbaren Reste in Prozent bezogen auf die Länge der jeweils kürzeren Proteinsequenz. Ein Vergleich der Balkendiagramme

FAM	#	SCOP Klasse	# Reste min	# Reste max	Mitglieder
4cyt	8	4-helical cytokines	119	172	1bgc, 1lki, 1huw, 1ilk, 1gmfA 1rcb, 1irl, 1rfbA, 1lpe
4hel	6	Four-helical up-and-down bundle	106	154	1was, 256bA, 2ccyA 2hmzA, 2tmvP, 1aep
OB	7	OB-FOLD	98	280	1ltsD, 1prtD, 1prtF, 1snc, 1pyp 2prd, 1gpc
cys	5	Cystine-knot cytokines	85	112	1pdgA, 2tgi, 1bndA, 1hcnA, 1hcnB
ferr	6	Ferredoxin-like	81	143	2fd2, 1pba, 1nhkL, 2bopA, 1aps, 1regX
flavo	5	Flavodoxin-like	128	302	3chy, 1rcf, 4fxn, 1cus, 1esc
glob	11	Globin-like	136	172	3sdhA, 2hbg, 2fal, 1hrm, 1eca, 2gdm 1pbxA, 1ash, 1hlb, 1cpcA, 1cpcB
hydro	6	α/β-Hydrolases	265	534	1ede, 1thtA, 1tca, 3tgl, 1crl, 1tahA
lipo	6	Lipocalins	131	176	1hbq, 1bbpA, 1epaB, 1mup, 1hmt
tim	11	α/β (TIM)-barrel	228	490	1byb, 1xyzA, 1pbgA, 1nar, 2ebn, 1fbaA 2acq, 1oyc, 1ubsA, 5timA, 1nfp
viral	10	Viral coat and capsid proteins	175	548	2bpa2, 2stv, 4sbvA, 1bbt1, 1bbt2, 1bbt3 4rhv3, 1pvc2, 2mev1, 2cas

Tabelle 8.5: Elf Faltungsfamilien, die nach SCOP [243] aus der Repräsentativmenge hobohm_96_25 [148] generiert wurden.

der verschiedenen Familien zeigt starke Unterschiede zwischen den strukturellen Ähnlichkeiten der Familienmitglieder untereinander. Während für die Vier–Helix–Bündel, die Cystein–Knoten Zytokine, die Flavodoxine, die Globine und die Lipocaline sehr häufig mehr als 50% der kürzeren Struktur mit einer *RMS*–Abweichung kleiner 3.4Å superpositioniert werden können, liegt dieser Wert in den restlichen Familien für viele der Paare weit niedriger. In diesen Familien existiert häufig sogar für eines oder mehrere Familienmitglieder kein Partner mit dem mehr als 50% der Reste gut superpositionierbar sind. Dieser Unterschied wird noch deutlicher, wenn, wie in Abbildung 8.16 gezeigt, die Anzahl der gut superpositionierbaren Reste auf die Resteanzahl der längeren Struktur bezogen wird. In der Familie der Vier–Helix–Zytokine fällt so zum Beispiel das γ–Interferon 1rfbA an jeweils letzter Position des *FAM-4cyt*–Teildiagramms auf. Bezogen auf die jeweils längere Sequenz sind zwischen 1rfbA und einem anderen Familienmitglied maximal 38% der Reste gut superpositionierbar. In der CATH–Klassifizierung (siehe Abschnitt 3.3.5.1) wird 1rfbA (CATH: 1 10 430) daher auch einer anderen Architektur–und Topologieklasse zugeordnet als die anderen Vier–Helix–Zytokine (CATH: 1 20 160). Eine Erkennung derart entfernter struktureller Ähnlichkeiten kann von Sequenzstrukturalignmentmethoden wie RDP kaum erwartet werden.

Die Mitglieder einer Familie unterscheiden sich teilweise durch ganze zusätzliche Sekundärstrukturelemente, deren Existenz die großen Längenunterschiede zwischen den Familienmitgliedern zur Folge haben. So hat zum Beispiel das Interleukin 10 1ilk zusätzlich zum Vier–Helix–Bündel zwei weitere Helices, die den anderen Familienmitgliedern fehlen. In den Familien der Vier–Helix–Bündel und der Cysteinknoten–Zytokine ist jedes Familienmitglied zumindest zu einem wei-

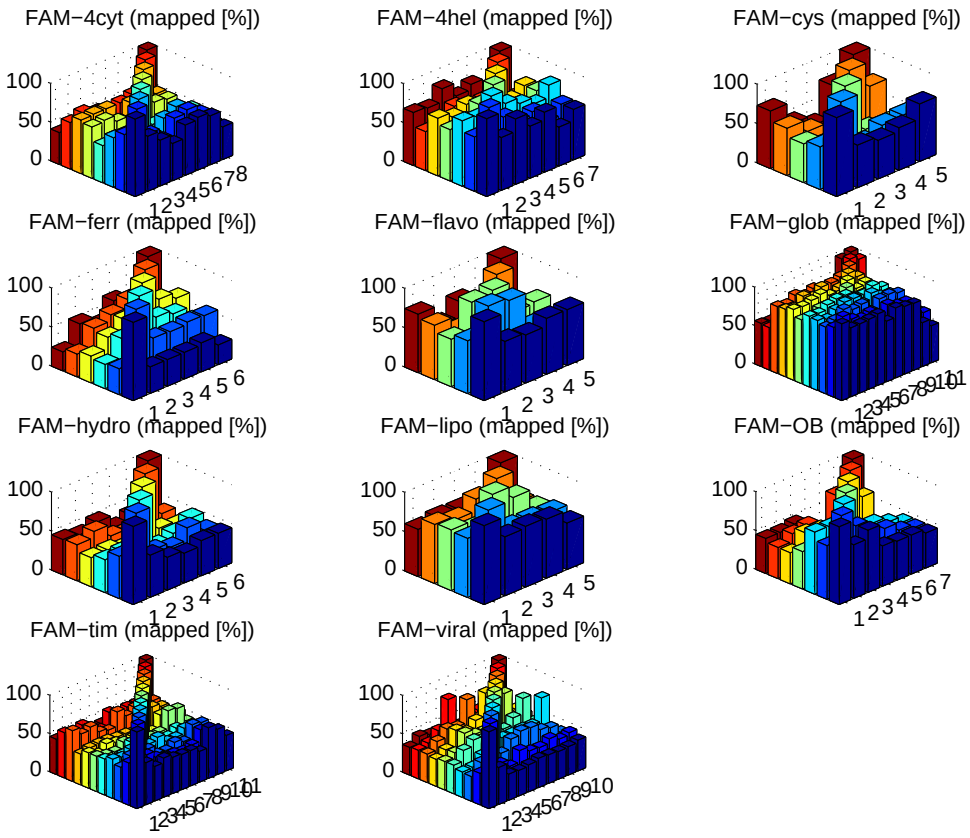


Abbildung 8.15: Prozentualer Anteil von `SARF2` mit einer maximalen *RMS*–Abweichung von 3.4Å superpositionierter Reste (bezogen auf die kürzere Sequenz).

terem Mitglied strukturell sehr ähnlich.
Dagegen weist die `SCOP`–Faltungsfamilie der Ferredoxin–ähnlichen Proteine eine große strukturelle Diversität auf. Sie umfaßt neben einem Ferredoxin (`2fd2`, `CATH 3 40 150`, Position 1), ein Regulatorprotein (`1regX`, `CATH 3 90 130`, Position 6) und vier Crambine (`CATH 3 30 70`). Abbildung 8.16 zeigt sehr deutlich, daß das Ferredoxin strukturell recht unterschiedlich (deutlich weniger als 30% superpositionierbar) zu den anderen Familienmitgliedern ist. Die Crambine sind untereinander am ähnlichsten, wobei zwischen 31 und 62% der Reste gut superpositionierbar sind.
Bei den Flavodoxin–ähnlichen Proteinen fällt die Esterase `1esc` sowohl aufgrund ihrer Länge (302 Reste gegenüber 128 bis 197 Reste sonst) und der abweichenden `CATH`–Klasse (`3 40 660` gegenüber `3 40 50` sonst) aus dem Rahmen. Die Esterase enthält jedoch das typische Faltungsmotiv der Flavodoxine. Daher superpositionieren zum Beispiel 74% der Reste von `3chy` gut mit den entsprechenden Resten aus `1esc`. Für Faltungserkennungsexperimente ist diese Art struktureller Ähnlichkeit nur schwer erkennbar, da ein strukturrichtiges Alignment eines Flavodoxins

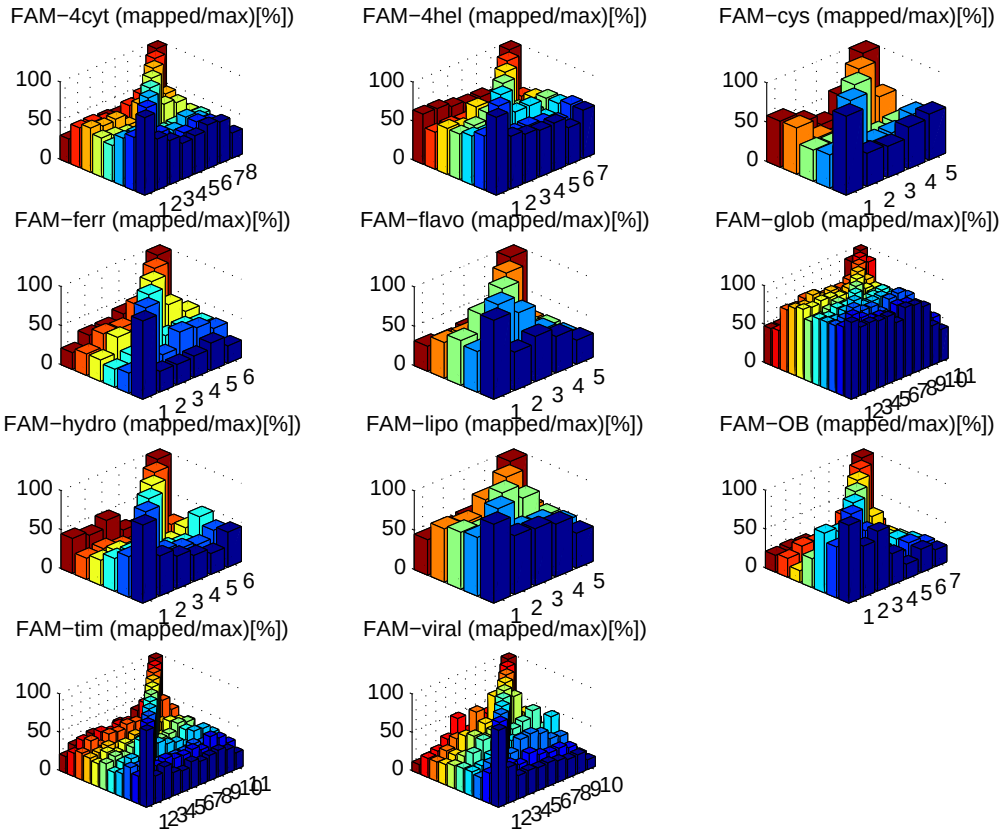


Abbildung 8.16: Prozentualer Anteil von `SARF2` mit einer maximalen *RMS*–Abweichung von 3.4Å superpositionierter Reste (bezogen auf die längere Sequenz).

mit der Esterase nur mit sehr vielen und sehr großen Insertionen beziehungsweise Deletionen möglich ist, was insbesondere in Faltungserkennungsexperimenten in aller Regel zu einer schlechten Bewertung des Paares führt.

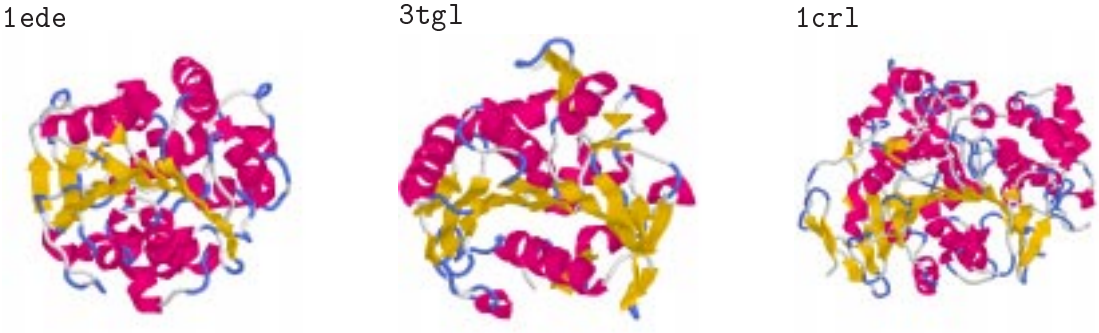


Abbildung 8.17: α/β-Hydrolasen: `1ede` als typische Lipase (`CATH 3 40 680`), die Nitrogenase `3tgl` (`CATH 3 40 50`) und die Acetylcholinesterase `1crl` (`CATH 3 40 680`)

Abbildung 8.17 zeigt die Unterschiedlichkeit der Faltungsfamilienmitglieder der α/β-Hydrolasen. Diese Familie beinhaltet vier Proteine die CATH als Lipasen (265 bis 318 Reste) klassifiziert, ein als Nitrogenase klassifiziertes Protein 3tgl (265 Reste) und das als Acetylcholinesterase klassifizierte Protein 1crl (534 Reste). Die Sonderrolle der Nitrogenase (Position 4) und der Acetylcholinesterase (Position 5) tritt in Abbildung 8.16 klar hervor, obwohl auch für die anderen Hydrolasen in allen Fällen weniger als 50% der Reste gut superpositionieren.
In der *OB*–Faltungsfamilie faßt SCOP im wesentlichen zwei Faltungstypen zusammen. Zum einen sind dies die Proteine (Positionen 1 bis 4), die nach CATH mit 2 40 50 als die eigentlichen *OB*–Faltungen klassifiziert werden und zwischen 98 und 135 Reste haben. Zum anderen fallen Proteine (Positionen 5 bis 7) in die *OB*–Faltungsfamilie, die mit 74 bis 280 Resten von CATH als 3 90 80/198) klassifiziert werden und neben dem zentralen, faßförmigen Faltblatt noch weitere Faltblätter und Helices haben. In Abbildung 8.15 sind diese zwei Unterklassen klar erkennbar. Faltungserkennungsmethoden können und sollten in diesem Fall natürlich jeweils nur ein Mitglied der zugehörigen Unterklasse erkennen, was die Auswahl der in der Testmenge zur Verfügung stehenden richtigen Partner stark einschränkt.
Warum die sogenannten *TIM–Barrel* für Faltungserkennungsmethoden im allgemeinen und für Methoden, deren zugrundeliegendes Bewertungssystem auf die Erhaltung von Paarinteraktionen abhebt, im besonderen eine sehr schwierige Faltungsfamilie darstellen, zeigt Abbildung 8.16 bereits auf den ersten Blick. Alle Familienmitglieder superpositionieren wechselseitig nur mit weniger als 50% ihrer Reste und sind zudem sehr unterschiedlich in ihren Längen (zwischen 228 und 490 Resten). Dabei wird die zentrale, faßförmige Faltblattstruktur durch unterschiedlich viele und an unterschiedlichen Stellen eingefügten Helices ergänzt, die dazu führen, daß in den verschiedenen Strukturen unterschiedliche Paarinteraktionen zu Resten in diesen zusätzlichen Faltungselementen ausgebildet werden.
Den größten Außenseiter bei den viralen Hüllproteinen bildet das Protein 2cas, das mit 548 Resten mehr als doppelt so lang ist, wie alle anderen Proteine der Familie (175 bis 268 Reste). Neben diesem Längenunterschied verdeutlicht Abbildung 8.16 aber auch, daß die Hüllproteine unterschiedlich gut wechselweise superpositionieren, so daß die recht hohe Anzahl von zehn Faltungsfamilienmitgliedern nur den Anschein erweckt, zu jedem Mitglied wären neun weitere Proteine mit gleicher Faltung in der Testmenge vorhanden.
Die obige Analyse der SCOP–Faltungsfamilien zeigt deutlich, daß es sich bei Erkennungsexperimenten auf dieser Testmenge um sehr schwierige Testbeispiele handelt. Dies liegt nicht nur daran, daß eine signifikante Sequenzähnlichkeit der Faltungsfamilienmitglieder untereinander durch die Wahl der Ausgangsmenge von vornherein ausgeschlossen ist, sondern auch daran, daß die strukturelle Ähnlichkeit der Familienmitglieder untereinander in sehr vielen Fällen recht gering ist. Die starken Unterschiede in der Sequenzlänge der Mitglieder der Faltungsfamilien erschweren die Erkennungsexperimente für alle Methoden, aber insbesondere

für die Methoden, bei denen Bestrafungsterme für Insertionen und Deletionen zentraler Bestandteil der Kostenfunktion sind.
Auf eine Bereinigung der Testmenge um die schwierigen, in einigen Fällen sehr wahrscheinlich sogar unlösbaren Beispiele wird verzichtet, um die Vergleichbarkeit gegen die bereits auf dieser Menge mit dem Programm `123D` durchgeführten Experimenten zu gewährleisten. Außerdem kann so festgestellt werden, wo für Sequenzstrukturalignmentmethoden die Grenzen der Erkennung entfernter struktureller Ähnlichkeiten liegen.

8.4.2 Bewertungskriterien für die Erkennungsexperimente

Das wichtigste Erfolgskriterium bei der Erkennung von entfernten Sequenzstrukturverwandtschaften ist, daß eine Proteinstruktur der gleichen Faltungsfamilie auf der ersten Position der Rangliste steht. Neben diesem Kriterium ($R1_{\mathcal{F}_i}$) werden in der folgenden Definition 8.1 einige weitere zur Analyse von Faltungserkennungsexperimenten benötigte Kriterien definiert:

Definition 8.1
Sei $B \in \mathcal{F}$ eine Faltung und $\mathcal{R}_B = [B_1, \ldots, B_n]$ die um die Struktur B bereinigte Rangliste eines Faltungserkennungslaufes mit der Sequenz B. Desweiteren seien $\mathcal{F}_i \subset \mathcal{F}$ Faltungsfamilien und

$$\delta_{\mathcal{F}_i}(B) = \begin{cases} 0 & , \textit{falls} \quad B \notin \mathcal{F}_i \\ 1 & , \textit{falls} \quad B \in \mathcal{F}_i \end{cases} .$$

Sei $B_1 = \mathcal{R}_B[1]$ die erste Struktur in der Rangliste $\mathcal{R}_B$, dann ist der Anteil $R1_{\mathcal{F}_i}$ der korrekt an Rang 1 erkannten homologen Strukturen einer Faltungsfamilie $\mathcal{F}_i$ wie folgt definiert:

$$R1_{\mathcal{F}_i} = \frac{\sum\limits_{B \in \mathcal{F}_i} \delta_{\mathcal{F}_i}(B_1)}{|\mathcal{F}_i|}$$

Der minimale Rang einer Struktur der Familie $\mathcal{F}_i$ in der Rangliste $\mathcal{R}_B$ wird durch die folgende Funktion bestimmt:

$$minrank(\mathcal{F}_i, \mathcal{R}_B) = \min_{\delta_{\mathcal{F}_i}(\mathcal{R}_B[j])=1} j$$

Das Maß $\bar{R}_{\mathcal{F}_i}$ mittelt den reziproken Rang der ersten homologen Faltung in den einzelnen Ranglisten einer Familie $\mathcal{F}_i$ über alle Ranglisten der Familie.

$$\bar{R}_{\mathcal{F}_i} = \frac{\sum\limits_{B \in \mathcal{F}_i} \frac{1}{minrank(\mathcal{F}_i, \mathcal{R}_B)}}{|\mathcal{F}_i|}$$

Sei $\phi(B, A)$ der Kostenfunktionswert oder Score *eines berechneten Sequenzstrukturalignments von Sequenz B in Struktur A. Dann ist $S_{\mathcal{F}_i}(B)$ ($S_{\neg\mathcal{F}_i}(B)$) der mittlere* Score *der Alignments von der Sequenz B in die Strukturen, die (*nicht*) zu*

der Familie $\mathcal{F}_i$ gehören, mit:

$$S_{\mathcal{F}_i}(B) = \frac{\sum_{A \in \mathcal{F}_i} \phi(B, A)}{|\mathcal{F}_i|}$$

$$S_{\neg\mathcal{F}_i}(B) = \frac{\sum_{A \notin \mathcal{F}_i} \phi(B, A)}{|\mathcal{F}| - |\mathcal{F}_i|}$$

Seien $\sigma_{\mathcal{F}_i}(B)$ und $\sigma_{\neg\mathcal{F}_i}(B)$ die zugehörigen Standardabweichungen (vgl. Definition 4.4). Dann definiert $Z_{\mathcal{F}_i}(B)$ einen familienspezifischen zscore von B:

$$Z_{\mathcal{F}_i}(B) = \frac{S_{\mathcal{F}_i}(B) - S_{\neg\mathcal{F}_i}(B)}{(\sigma_{\mathcal{F}_i}(B) + \sigma_{\neg\mathcal{F}_i}(B))/2}$$

Der normierte Mittelwert $\bar{Z}_{\mathcal{F}_i}$ des familienspezifischen $Z_{\mathcal{F}_i}$ ist dann:

$$\bar{Z}_{\mathcal{F}_i} = \frac{\sum\limits_{B \in \mathcal{F}_i} Z_{\mathcal{F}_i}(B)}{|\mathcal{F}_i|}$$

Ein $R1_{\mathcal{F}_i}$ von 1.0 bedeutet, daß für *alle* Erkennungsläufe für Familie $\mathcal{F}_i$ eine homologe Struktur auf Ranglistenposition 1 war. Ein Wert von 0.5 bedeutet, daß in jedem zweiten Fall eine homologe Struktur auf Position 1 gefunden wurde.
Dagegen mißt $\bar{R}_{\mathcal{F}_i}$ den mittleren Rang, an dem für ein Protein aus der Familie $\mathcal{F}_i$ das erste Familienmitglied in der Rangliste gefunden wurde. Ein Wert von 0.5 bedeutet also, daß im Mittel über alle Familienmitglieder eine homologe Struktur auf Position 2 der Rangliste gefunden wird. Je höher also $\bar{R}_{\mathcal{F}_i}$ ist, desto niedriger ist der Rang der ersten homologen Struktur in der Rangliste.
Der normierte Mittelwert $\bar{Z}_{\mathcal{F}_i}$ des familienspezifischen *zscores* dient als Maß für den Unterschied in der Bewertung von Paaren, die aus Mitgliedern der gleichen Familie bestehen, gegenüber Paaren, die aus unterschiedlichen Faltungsfamilien stammen.

8.4.3 Ergebnisse der Erkennungsexperimente

In den vorangegangenen Abschnitten sind die Testmenge (siehe Abschnitt 8.4.1) und die Bewertungskriterien (siehe Abschnitt 8.4.2) für die im folgenden durchgeführten Erkennungsexperimente beschrieben worden. Zur Bewertung der Erkennungsleistung der RDP–Methode werden 81 unabhängige Erkennungsexperimente (siehe auch Definition 4.2 und Abbildung 4.5) durchgeführt. Für jedes Element der oben beschriebenen Faltungsfamilien wird in der Menge der 251 Proteine mit einer Domäne aus `hobohm97_25` nach möglichen Faltungen gesucht. Ein Erkennungsexperiment war dann erfolgreich, wenn ein Mitglied der Faltungfamilie, zu der das untersuchte Protein gehört, auf Position 1 der Rangliste der 251 Sequenzstrukturalignments gefunden wird ($R1$–Kriterium aus Definition 8.1).

Da bei diesem Test die native Struktur selbstverständlich ausgeschlossen wird, bedeutet ein Erfolg aufgrund der Wahl der Testmenge die Erkennung einer verwandten Struktur, deren Sequenzidentität zum untersuchten Protein geringer als 25% ist.
Um die mit der RDP–Methode erzielten Ergebnisse einzuordnen, werden die Erkennungsexperimente ebenfalls mit Sequenzalignmentmethoden [226] und dem Faltungserkennungsprogramm `123D` [5] durchgeführt. Ein Vergleich mit dem Programm `Threader` ist zum jetzigen Zeitpunkt aus Laufzeitgründen leider nicht möglich, da `Threader` für einen Vergleich die 81 zu den Faltungsfamilien gehörigen Sequenzen jeweils gegen 251 Strukturen zu alinieren hätte. Daher bleibt hier nur der Vergleich gegen das Sequenzalignment und gegen die schnelle `123D`–Methode, die zudem beim Vergleich der Methoden hinsichtlich der Alignmentqualität weitaus bessere Ergebnisse gezeigt hat, als dies sowohl für `Threader1` als auch `Threader2` der Fall war (siehe Abschnitt 8.3).

Sequenzalignment		$\bar{Z}$			$R1$			$\bar{R}$		
Familie	#	day	gon	b62	day	gon	b62	day	gon	b62
FAM-4cyt	8	1.00	1.02	0.99	0.25	0.50	0.50	0.57	0.69	0.70
FAM-4hel	7	1.10	1.10	1.08	0.17	0.33	0.33	0.30	0.37	0.39
FAM-OB	7	0.64	0.62	0.66	0.14	0.14	0.14	0.18	0.18	0.19
FAM-cys	5	1.16	1.18	1.16	0.00	0.00	0.00	0.20	0.10	0.11
FAM-ferr	6	1.14	1.15	1.14	0.00	0.00	0.17	0.14	0.19	0.27
FAM-flavo	5	0.76	0.75	0.77	0.00	0.20	0.20	0.16	0.28	0.32
FAM-tim	11	0.86	0.87	0.78	0.09	0.09	0.09	0.29	0.22	0.23
FAM-glob	11	1.33	1.33	1.23	1.00	1.00	1.00	1.00	1.00	1.00
FAM-hydro	6	1.04	1.05	0.97	0.00	0.17	0.00	0.19	0.30	0.19
FAM-lipo	5	1.21	1.23	1.18	0.80	0.80	0.80	0.80	0.80	0.80
FAM-viral	10	0.66	0.66	0.63	0.50	0.40	0.40	0.57	0.46	0.49
Mittelwert		0.99	1.00	0.96	0.27	0.33	0.33	0.40	0.42	0.43

Tabelle 8.6: Faltungserkennung mittels Sequenzalignment [226]: Die Spalten `day`, `gon` und `b62` zeigen die Ergebnisse für die verwendete Austauschmatrix (`dayhoff`, `gonnet` beziehungsweise `blosum62`, siehe 5.1). Als Gapkosten wurden einheitlich als Gaperöffnungskosten 15.0 und 3.0 als Gapverlängerungskosten verwendet.

Tabelle 8.6 zeigt die Erkennungsleistung, die mit reinen Sequenzalignmentmethoden [226] bei Verwendung verschiedener Austauschmatrizen auf dieser Testmenge erreichbar ist. Ein Vergleich der zu den Austauschmatrizen gehörigen Spalten zeigt, daß modernere Matrizen wie `gonnet` und `blosum62` nicht nur — wie in der Literatur vielfach nachgewiesen (siehe dazu auch Abschnitt 5.1) — zu einer höheren Alignmentqualität, sondern auch zu einer Steigerung der Erkennungsleistung

von Sequenzalignmentmethoden führen. Wie die Mittelwertzeile der Tabelle 8.6 zeigt, gelingt es auch mit den modernen Austauschmatrizen, nur in einem Drittel der Beispiele eine verwandte Struktur auf Platz 1 der Rangliste zu finden. Angesichts der Tatsache, daß für alle hier betrachteten Beispiele die Sequenzidentität weit unter 25% liegt, ist diese Erkennungsleistung durchaus bemerkenswert.

`123D` CCP+ ...		$\bar{Z}$			$R1$			$\bar{R}$		
Familie	#	day	gon		day	gon		day	gon	
FAM-4cyt	8	1.11	1.13	1.24	0.12	0.12	0.12	0.32	0.35	0.23
FAM-4hel	7	1.38	1.40	1.62	0.33	0.33	0.17	0.40	0.41	0.29
FAM-OB	7	0.71	0.70	0.54	0.14	0.14	0.14	0.28	0.27	0.26
FAM-cys	5	1.15	1.16	1.19	0.20	0.20	0.20	0.36	0.36	0.29
FAM-ferr	6	1.04	1.03	1.11	0.17	0.17	0.33	0.26	0.29	0.37
FAM-flavo	5	1.11	1.10	1.08	0.40	0.60	0.60	0.50	0.44	0.39
FAM-tim	11	1.19	1.24	1.75	0.45	0.45	0.45	0.66	0.65	0.62
FAM-glob	11	1.41	1.41	1.53	1.00	1.00	0.45	1.00	1.00	0.57
FAM-hydro	6	1.19	1.22	1.66	0.17	0.00	0.00	0.33	0.27	0.26
FAM-lipo	5	1.48	1.51	1.70	0.60	0.60	0.60	0.74	0.72	0.71
FAM-viral	10	0.88	0.90	0.99	0.60	0.60	0.60	0.44	0.44	0.41
Mittelwert		1.15	1.16	1.31	0.38	0.38	0.33	0.48	0.47	0.40

Tabelle 8.7: Faltungserkennung mit der 123D–Methode [379]: Für die Spalten `day` und `gon` wurde jeweils zusätzlich zum `CCP`–Potential eine Austauschmatrix (`dayhoff` beziehungsweise `gonnet`, siehe 5.1) als Bestandteil der Bewertungsfunktion verwendet.

Tabelle 8.7 zeigt die Faltungserkennungsergebnisse, die mit der `123D`–Methode [5] auf der beschriebenen Testmenge erzielt werden. Es wird unterschieden, ob und wenn ja welche Aminosäureaustauschmatrix zusätzlich neben dem Kontaktkapazitätspotential (`CCP`) für die Berechnung und Bewertung der Sequenzstrukturalignments verwendet wurde. Ein Vergleich der `dayhoff`– und `gonnet`–Spalten gegen die Spalten, in denen das Kontaktkapazitätspotential allein verwendet wurde, zeigt, daß auch für Proteine sehr niedriger Sequenzidentität die Sequenzähnlichkeit einen wesentlichen Beitrag zur Erkennung leisten kann. Bis auf die Faltungsfamilie der Ferredoxine verbessert die Hinzunahme der Sequenzähnlichkeit in allen Familien die Erkennungsleistung der `123D`–Methode, sowohl was die Rang 1 Erkennung ($R1$) als auch was den mittleren Rang des ersten Familienmitglieds in der Rangliste ($\bar{R}$) betrifft.
Wie der Vergleich des Mittelwertes des familienspezifischen *zscores* $\bar{Z}$ zeigt, scheint durch die Hinzunahme von Sequenzinformation die Trennung zwischen den zu der gleichen Familie gehörenden und damit korrekten Treffern in der Rangliste und den nicht zu der gleichen Familie gehörenden Proteinen verwischt zu werden. Im

Durchschnitt über alle Familien der Testmenge sinkt der familienspezifische *zscore* $\bar{Z}$ von 1.3 ohne Sequenzanteil auf 1.15 mit Austauschmatrix ab – unabhängig davon, ob `dayhoff` oder `gonnet` als Austauschmatrix verwendet wird.
Ein Vergleich der für die einzelnen Familien erzielten Erkennungsraten bestätigt die bereits bei der Analyse der Testbeispiele aufgestellte Hypothese, daß die Erkennung strukturell ähnlicher Faltungen an Platz 1 der Rangliste ($R1$–Maß) für einige der Familien leichter und für andere schwerer ist. Am besten ist die Erkennung für die Faltungsfamilie der Globine. Mit Sequenzinformation findet `123D` für alle Globine eine homologe Faltung auf Rang 1 der Rangliste. Bei den Familien der Flavodoxin–ähnlichen Proteine, der Proteine mit *TIM–barrel*–Faltung, der Lipocaline und der viralen Hüllproteine werden für 45 bis 60% der Beispiele ähnliche Faltungen auf Rang 1 erkannt.
Für die Familien der Vier–Helix–Zytokine, der Vier–Helix–Bündel, der OB–Faltungen, der Cystein–Knoten–Zytokine, der Ferredoxin–ähnlichen Proteine und der Hydrolasen befindet sich für weniger als ein Drittel der Familienmitglieder ein strukturell verwandtes Protein auf Rang 1 der `123D`–Rangliste. Bis auf die Familie der Vier–Helix–Bündel handelt es sich dabei um die Familien, die bereits in Abschnitt 8.4.1 als schwierig und problematisch identifiziert wurden. Warum die Erkennung der `123D`–Methode für die Faltungsfamilie der Vier–Helix–Bündel nur 17% beträgt, ist unklar und sollte genauer untersucht werden. Das $\bar{R}$–Maß zeigt jedoch an, daß die strukturähnlichen Proteine auch für diese Familie im oberen Bereich der Rangliste zu finden sind.
Wie die Mittelwertzeile von Tabelle 8.7 zeigt, findet die `123D`–Methode in 38% von 81 Fällen ein Protein der gleichen Familie auf Rang 1 der Rangliste mit 251 Proteinstrukturen. Das erste strukturähnliche Protein findet sich im Mittel knapp unterhalb der Position 2 der Rangliste ($\bar{R} = 0.48$ für `CCP` mit `dayhoff`–Matrix). Die Bewertung eines Paares aus Mitgliedern der gleichen Familie ist im Mittel um mehr als eine Standardabweichung ($\bar{Z} \geq 1.15$ für alle drei Spalten) höher als die von Paaren, die aus strukturell unähnlichen Proteinen bestehen.
Die Tabelle 8.8 zeigt die mit der RDP–Methode erzielten Erkennungsergebnisse. Für die Faltungserkennungsexperimente mit RDP werden die in Abschnitt 7.7 empirisch bestimmten Parameter verwendet. Zusätzlich zu dem auf Voronoikontaktrelationen basierenden Potential (`VCM`–Spalten) wird zum Vergleich auch ein auf Distanzkontaktrelationen basierendes Potential (PDB–Spalten) unter Beibehaltung der weiteren Parameter getestet. Das auf Distanzkontaktrelationen basierende Potential erkennt in 47% aller Beispiele ein strukturähnliches Protein auf Position 1 der Rangliste. Dies sind 9% mehr als mit `123D` und sogar 14% mehr als mit Sequenzalignment erkannt werden (vergleiche Tabellen 8.7 und 8.6).
Bei Verwendung des auf Voronoikontaktrelationen basierenden Potentials werden sogar in 53% der Beispiele Proteine der gleichen Faltungsfamilie auf Rang 1 der Rangliste gefunden. Dies sind 15(20)% mehr als mit `123D` (Sequenzalignment) zu finden sind. Doch nicht nur die $R1$–Erkennungsrate kann durch Verwendung der RDP–Methode gesteigert werden, sondern die RDP–Methode findet im Durch-

RDP		$\bar{Z}$		$R1$		$\bar{R}$	
Familie	#	PDB	VCM	PDB	VCM	PDB	VCM
FAM-4cyt	8	-0.89	-0.87	0.25	0.38	0.48	0.51
FAM-4hel	7	-1.04	-1.04	0.57	0.71	0.66	0.72
FAM-OB	7	-0.86	-0.79	0.57	0.29	0.61	0.42
FAM-cys	5	-1.15	-1.17	0.20	0.40	0.33	0.46
FAM-ferr	6	-0.92	-0.87	0.17	0.17	0.25	0.22
FAM-flavo	5	-1.24	-1.09	0.20	0.60	0.43	0.62
FAM-tim	11	-0.82	-0.88	0.18	0.27	0.33	0.44
FAM-glob	11	-1.65	-1.51	1.00	1.00	1.00	1.00
FAM-hydro	6	-1.03	-1.08	0.17	0.50	0.41	0.61
FAM-lipo	5	-1.37	-1.44	0.80	0.80	0.81	0.90
FAM-viral	10	-0.83	-0.91	0.70	0.60	0.72	0.69
Mittelwert		-1.07	-1.05	0.47	0.53	0.57	0.61

Tabelle 8.8: Faltungserkennung mit der RDP–Methode: Für die VCM–Spalten wurde ein auf Voronoikontaktrelationen basierendes Potential und für die PDB–Spalten ein auf Distanzkontaktrelationen basierendes Potential verwendet.

schnitt auch ein Familienmitglied weiter oben in der Rangliste als dies bei `123D` der Fall ist, wie ein Vergleich der $\bar{R}$–Spalten in Tabellen 8.8 und 8.7 zeigt.
Der familienspezifische *zscore* $\bar{Z}$ ist bei der RDP–Methode mit einem negativen Vorzeichen versehen, da RDP im Unterschied zum Sequenzalignment und zu `123D` minimiert statt maximiert. Die Trennung der Familienmitglieder von den nicht Familienmitgliedern bezüglich der Kostenfunktionsbewertung gemessen in Einheiten Standardabweichung ist bei der RDP–Methode etwas geringer als bei `123D`. Ein Grund dafür sind sicher die geringen Kosten für Insertionen und Deletionen in der Kostenfunktion der RDP–Methode. Die niedrigen Gapkosten machen es möglich, daß zum Beispiel beim Alignment einer kürzeren Sequenz gegen eine längere Faltung die gut passenden Bereiche für das Alignment selektiert werden, die dann in der Regel aus unzusammenhängenden Sekundärstrukturelementen bestehen. Wie die $R1$–Erkennung zeigt, hat die Vorgehensweise, auch möglichst gute Alignments zwischen nicht zu einer Familie gehörigen Sequenzstrukturpaaren zu berechnen, keinen negativen Einfluß auf die Erkennung wirklich ähnlicher Faltungen an Platz 1 der Ranglisten.
Wie auch bei der `123D` stellen sich auch bei der RDP–Methode die Familien als schwer heraus, die bereits bei der Voranalyse der Testmenge als schwer eingestuft wurden.
Wesentliche Verbesserungen gegenüber `123D` kann die RDP–Methode für die Proteine mit Vier–Helix–Bündel–Faltung (71% *vs.* 33%), die Vier–helikalen Zytokinen (38% *vs.* 12%) und die Hydrolasen (50% *vs.* 17%) erreichen. Kleiner fallen

die Verbesserungen für die OB–Faltungen (29% *vs.* 14%), die Cystein–Knoten Zytokine (40% *vs.* 20%) und die Lipocaline (80% *vs.* 60%) aus. Bei den OB–Faltungen erkennt RDP mit dem auf Distanzkontaktrelationen basierenden Potential sogar in 57% (`123D` 14%) ein Element der Familie auf Rang 1. Allein bei den *TIM–barrel*–Faltungen erkennt RDP mit 27% weniger Familienmitglieder als `123D` (45%) auf Rang 1.
Wie die Analyse der Alignmentqualität (Abschnitt 8.3) gezeigt hat, berechnet die RDP–Methode sehr gute Sequenzstrukturalignments. Die Qualität der Methode zeigt sich insbesondere dann, wenn Insertionen und Deletionen nur gering bestraft werden, so daß wirklich nur die zueinander passenden Bereiche aliniert werden. Die strukturrichtigen Alignments von ähnlichen Faltungen unterscheiden sich jedoch von denen für unähnliche Faltungen sehr häufig insbesondere durch die beim strukturrichtigen Alignment notwendigen Insertionen und Deletionen. Daher kann eine etwas höhere Gewichtung der Gapkosten für die Trennung der Alignments ähnlicher und unähnlicher Faltungen von Nutzen sein.
Im folgenden wird ein erster Versuch unternommen, die Erkennungsleistung allein durch eine Neubewertung der bereits berechneten Alignments zu erhöhen. Im Detail werden sogar alle Kostenfunktionsbestandteile, die bei der Berechnung der Alignments verwendet wurden (wie zum Beispiel die Potentiale), beibehalten und es wird versucht, nur durch Einstellung der Gewichtungsfaktoren der Kostenfunktionsbestandteile untereinander (siehe Abschnitt 7.7) die Erkennungsleistung zu verbessern.
Ziel der Umgewichtung der Kostenfunktionsbestandteile ist, in möglichst vielen der 81 Ranglisten ein Mitglied der Faltungsfamilie auf Platz 1 zu bekommen, zu der auch die jeweils untersuchte Proteinsequenz gehört. Das aus dieser Problembeschreibung resultierende Optimierungsproblem kann als quadratisches Zuordnungsproblem [198] formuliert werden. Da quadratische Zuordnungsprobleme *NP*–schwer [108] sind, wird hier eine von Alexander Zien [378] für die Optimierung allgemeiner Parameterprobleme entwickelte Methode zur Berechnung heuristischer Lösungen verwendet. In dieser Methode wird die Anzahl der in einem linearen Ungleichungssystem verletzten Ungleichungen minimiert. Dabei wird für jedes Paar aus der Rangliste eine Ungleichung aufgestellt, die beschreibt, welches der beiden Proteine strukturell ähnlicher zu der untersuchten Sequenz ist.
Da die Methode darauf abzielt, die Anzahl der durch die Rangliste verletzten Ungleichungen zu minimieren, wird durch sie die Erkennungsleistung nur indirekt verbessert. Aufgrund des dieser Methode zugrunde gelegten Optimierungskriteriums ist es attraktiver, eine Struktur aus der jeweils richtigen Faltungsfamilie, die sich aufgrund eines schlechten Sequenzstrukturalignments auf Rang 200 befindet, auf Rang 190 hoch zu bringen, als eine andere ebenfalls richtige Struktur von Rang 3 auf 1 umzusortieren, zu der zusätzlich ein wesentlich besseres Sequenzstrukturalignment berechnet werden konnte. Dies liegt daran, daß im ersten Fall neun und im zweiten Fall nur zwei weitere Ungleichungen durch die Umgewichtung der Parameter erfüllt werden. Bei Erkennungsexperimenten kommt es aber darauf an,

eines der Familienmitglieder auf Rang 1 und möglichst viele Familienmitglieder auf den ersten Positionen der Rangliste zu plazieren. Auf welchem Rang Familienmitglieder mit einem schlechten Sequenzstrukturalignment stehen, interessiert dagegen in der Regel wenig, solange auf den oberen Plätzen der Rangliste die richtigen Strukturvorschläge zu finden sind.

RDPrerank		$\bar{Z}_{\mathcal{F}_i}$		$R1$		$\bar{R}$	
Familie	#	PDB	VCM	PDB	VCM	PDB	VCM
FAM-4cyt	8	-0.98	-0.88	0.50	0.25	0.68	0.45
FAM-4hel	7	-1.19	-1.02	0.71	0.71	0.72	0.74
FAM-OB	7	-0.88	-0.83	0.57	0.57	0.61	0.63
FAM-cys	5	-1.27	-1.16	0.20	0.60	0.36	0.64
FAM-ferr	6	-1.00	-0.89	0.00	0.17	0.10	0.22
FAM-flavo	5	-1.22	-1.05	0.60	0.60	0.68	0.64
FAM-tim	11	-0.74	-0.77	0.27	0.36	0.47	0.60
FAM-glob	11	-1.77	-1.35	1.00	1.00	1.00	1.00
FAM-hydro	6	-1.01	-0.99	0.17	0.50	0.33	0.54
FAM-lipo	5	-1.54	-1.33	1.00	0.80	1.00	0.90
FAM-viral	10	-0.95	-0.88	0.70	0.60	0.78	0.69
Mittelwert		-1.13	-1.00	0.54	0.57	0.64	0.66

Tabelle 8.9: Faltungserkennung auf von der RDP–Methode mit den für Tabelle 8.8 verwendeten Parametern berechneten Sequenzstrukturalignments und einer nachträglichen Neubewertung der Alignments.

Aus diesen Gründen wurde die Parameterkalibrierung im wesentlichen experimentell eingesetzt, um verschiedene Parameterkombinationen auf ihre Erkennungsleistung hin zu testen, ohne die Sequenzstrukturalignments aufwendig neu berechnen zu müssen. Tabelle 8.9 zeigt die besten bei diesem Test erzielten Erkennungsergebnisse, die durch eine Kombination manuell vorgenommener Parametereinstellungen und Kalibrierung erreicht wurden.
Wie ein Vergleich von Tabelle 8.9 mit Tabelle 8.8 zeigt, kann durch eine Verschiebung der Gewichtungsparameter eine sieben– beziehungsweise vierprozentige Verbesserung der $R1$–Erkennung erreicht werden. Die bisherigen Versuche bestätigen dabei die Vermutung, daß eine Verbesserung der Erkennungsleistung im wesentlichen, über die höhere Bestrafung von Insertionen und Deletionen möglich ist, während die Gewichtung der anderen Kostenfunktionsterme untereinander im wesentlichen erhalten bleibt. Eine entsprechend hohe Bestrafung von Insertion und Deletionen bereits bei der Berechnung der Sequenzstrukturalignments hat sich jedoch als nachteilig erwiesen, da dann die Alignmentqualität abnimmt.
Abbildung 8.18 faßt noch einmal die diskutierten Faltungserkennungsergebnisse auf einem Blick zusammen. Für alle Familien mit Ausnahme der *TIM–barrel–*

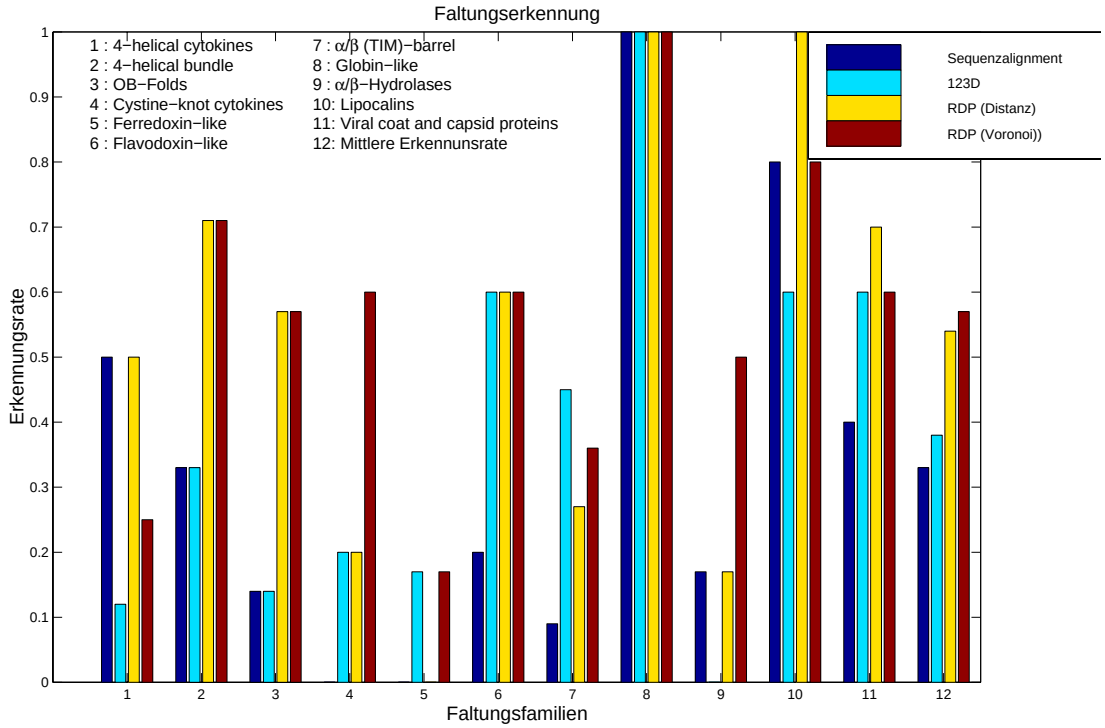


Abbildung 8.18: Vergleich der Erkennungsraten der verschiedenen Methoden: Sequenzalignment (blau), `123D` (cyan), RDP mit Distanzkontaktrelationen (gelb) und Voronoikontaktrelationen (rot).

Faltungen erkennt die RDP–Methode häufiger eine ähnliche Faltung als die anderen Methoden. In allen Fällen mit Ausnahme der Lipocaline und der viralen Hüllproteine werden mit den auf Voronoikontaktrelationen basierenden Kontaktpotentialen mehr richtige Faltungen erkannt als mit dem auf Distanzkontaktrelationen basierenden Potential.

Mit der RDP–Methode kann also nicht nur die Alignmentqualität, sondern über die qualitativ guten Sequenzstrukturalignments auch die Erkennungsrate gegenüber reinem Sequenzalignment (siehe Tabelle 8.6) und gegenüber `123D` (siehe Tabelle 8.7) beträchtlich gesteigert werden. Diese Steigerung beträgt auf der hier verwendeten Testmenge, deren Schwierigkeiten in Abschnitt 8.4.1 veranschaulicht wurden, insgesamt bei der Verwendung von auf Voronoikontaktrelationen basierenden Potentialen 24% gegenüber dem Sequenzalignment und immerhin noch 19% gegenüber `123D`. Die zwölfte Spalte der Abbildung 8.18 zeigt den direkten Vergleich der Erkennungsraten gemittelt über alle Familien der Testmenge. Daß die $R1$–Erkennungsrate mit maximal 57% immer noch nicht optimal ist, liegt in der Schwierigkeit der verwendeten Testbeispiele begründet.

8.5 Strukturvorhersagewettbewerb: CASP II

Die RDP–Methode wurde wie auch die anderen `ToPLign`–Methoden, Sequenzalignment und `123D` im Rahmen der Strukturvorhersagen für den `CASP II` Wettbewerb im Zeitraum Juni bis Oktober 1996 eingesetzt. Schon die Teilnahme an dem Wettbewerb zeigte Verbesserungsmöglichkeiten sowohl in der Erkennungsmethode und den Kostenfunktionen als auch im Funktionsumfang und in der Bedienerfreundlichkeit der Programme auf, die in die Weiterentwicklung insbesondere der RDP–Methode eingeflossen sind. Neben der Erprobung der eigentlichen Faltungserkennungsprogramme `123D` und RDP wurden schon während des Vorhersagezeitraums zahlreiche Programme zur Nachanalyse berechneter Alignments entwickelt, die zum Beispiel Bewertungen mit verschiedenen Potentialen, positionelle Analyse oder graphische Ergebnisaufbereitungen durchführen oder den Vorhersageprozeß in Teilen automatisieren.
Bei den einzelnen Vorhersagen im Rahmen von `CASP II` wurde versucht, die Ergebnisse der unterschiedlichen Verfahren durch eine konsistente Vorhersage in Übereinstimmung zu bringen. Dieser Prozeß wurde weitgehend manuell durchgeführt, obwohl die entwickelten Verfahren zum Vergleich von Alignments und der Berechnung von Alignmentdistanzen eingesetzt wurden. Die durchgeführten Vergleiche haben Schwachstellen oder Grenzen der jeweiligen Methoden beziehungsweise der Bewertungs– und Ranglistenkriterien aufgedeckt. Die durch den Vergleich gewonnenen Erkenntnisse wurden neben der Konstruktion einer konsistenten Vorhersage auch zur gegenseitigen Kalibrierung der Verfahren, Parameter und Kostenfunktionen eingesetzt.
Im Zeitraum von Juni bis Oktober 1996 waren insgesamt 22 Proteinstrukturen für unterschiedlich lange Sequenzen aus verschiedenen Organismen mit unterschiedlichen biologischen Funktionen und einer ganzen Bandbreite von Strukturklassen vorherzusagen. Die Ergebnisse sind in den Tabellen 8.10– 8.13 zusammengefaßt. Für jede der Vorhersagen gab es kurze Vorhersagezeiträume mit definierten Abgabedaten. Die eingereichten Strukturvorhersagen wurden von einem unabhängigen Auswertekommitee registriert und schließlich für die Auswertung beim Evaluierungsworkshop im Dezember vorbereitet. Von den 22 vorhergesagten Proteinen wurden nur 14 rechtzeitig zur Auswertung experimentell aufgeklärt, bei 8 Proteinen konnten die Experimentatoren den prognostizierten Zeitplan der Strukturaufklärung nicht einhalten.

Target-Id.	# Reste	Beschreibung	Vorhersagen		NONE
t23	284	KDO8P-Synthase,E.coli	1GOX- 0.25	1WSYA 0.25	0.5
t19	340	RepA1,E.coli	1CGPA 0.1	-	0.9

Tabelle 8.10: `CASP II`: eingreichte aber nicht ausgewertete Targets [200].

Von den 22 Targets wurden an der GMD 16 bearbeitet und eingereicht, zwei waren für ähnlichkeitsbasierte Methoden ungeeignet, eine Vorhersage wurde nicht

rechtzeitig fertiggestellt, und für eine waren die Ergebnisse inkonsistent (siehe Tabelle 8.11), zwei der Targets wurden nicht rechtzeitig experimentell aufgeklärt und kurzfristig aus dem Wettbewerb genommen (siehe Tabelle 8.10).

Target-Id.	# Reste	Beschreibung	Grund
t6	269	Outer-Membrane-Phospholipase-A, E.coli	noch offen
t8	29	De-Novo-designed-peptide	zu kurz für Threading
t11	220	Hsp-90, N-terminal-domain, S. cerevisiae	Vorhersage nicht rechtzeitig fertig
t21	64	KorB, C-terminal-domain, E.coli	noch offen
t30	66	domain-1, protein-g3, filamentous-phage-fd	zu kurz für Threading
t37	109	calponon-homology-domain, β-spectrin, Human	schlecht interpretierbare Ergebnisse

Tabelle 8.11: `CASP II`: nicht eingereichte Targets [200].

Für jede der verbleibenden 16 Vorhersagen wurden neben der in dieser Arbeit beschriebenen RDP–Methode auch die anderen im `PROTAL`–Projekt entwickelten Methoden eingesetzt. Da die Wahl der Parameter sowohl für die Faltungserkennung und als auch das genaue Sequenzstrukturalignment entscheidend ist, wurden für jedes Target und jede Methode verschiedene plausible Parameterkombinationen und unterschiedliche Repräsentativmengen durchgerechnet.
Im ersten Schritt der Vorhersage wurden für jedes Vorhersagetarget die bekannten Fakten aus der relevanten biologischen Literatur zusammengetragen und Hinweise und Restriktionen für entsprechende Strukturvorschläge abgeleitet. Parallel wurden alle an der GMD entwickelten Methoden zur Suche nach zu den Vorhersagetargets homologen Strukturen eingesetzt und durch schrittweise Verfeinerung die Menge der möglichen Kandidaten eingeschränkt. Danach wurden die mit den verschiedenen Methoden und Parametersätzen abgeleiteten Vorhersagen miteinander verglichen und zu einem konsistenten Strukturvorschlag verdichtet, der möglichst mit den biologischen Fakten in Einklang stand. Schließlich wurden die plausiblen Strukturvorschläge mit Wahrscheinlichkeiten versehen und zu einem detaillierten Sequenzstrukturalignment verfeinert. Dazu wurden neben den beiden Sequenzstrukturalignmentmethoden `123D` und RDP vor allem Analysen der Kompaktheit der Alignments, Vergleiche mit der vorhergesagten oder aus der Literatur bekannten Sekundärstruktur, sowie Berechnungen und Korrelationen positioneller Paarpotentialverläufe verwendet.
Im Dezember 1996 fand in Asilomar, Kalifornien, der Auswerteworkshop statt. Dabei wurden Kriterien zur Bewertung von Strukturvorhersageergebnissen und für den Vergleich unterschiedlicher Vorhersagemethoden diskutiert und auf die eingereichten Vorhersagen angewendet. Wie bei einem solchen hochkompetitiven Wettbewerb zu erwarten, gab es ausgiebige Diskussionen und Meinungsverschiedenheiten unter den Experten. Insgesamt kann man aber festhalten, daß

von den vier Wettbewerbskategorien „*Ab–Initio–Prediction*“, „*Comparative modelling*“, „*Fold Recognition and Threading*“ und „*Docking*“, am für unsere Einreichungen relevanten Teil „*Fold Recognition and Threading*“ die meisten Arbeitsgruppen teilnahmen und hier auch der größte Fortschritt gegenüber dem letzten *state–of–the–art* Workshop `CASP I` (1994) zu verzeichnen war. Konsequenterweise gab es hier auch die meisten Diskussionen um die Bewertung der eingereichten Vorhersagen und die dazu eingesetzten Methoden.
Die harte Konkurrenz im *Threading*–Bereich demonstriert auch die allgemein erwartete große Bedeutung solcher Methoden im Rahmen der Nutzung der in den Genomprojekten massiv anfallenden Sequenzdaten für industrielle Forschungs– und Anwendungsprojekte.
Die Tabelle 8.12 gibt eine Übersicht über unsere Vorhersagen für jene Targets in der Kategorie *Threading*, bei denen die gemessene Struktur sich als hinreichend ähnlich zu einem der Proteine in der Datenbank herausgestellt hat. In der Tabelle werden unsere Vorhersagen in den drei Spalten `VAST`, `DALI` und `SSAP` auf der Basis der mit den bei der Ergebnisbewertung eingesetzten Strukturvergleichsmethoden erzielten Ergebnisse bewertet. In fünf weiteren Spalten ist die Übereinstimmung der Vorhersage mit der experimentell ermittelten Struktur hinsichtlich der Klasse, der Faltung, der Topologie, der Struktur und dem Alignment angegeben. Bezüglich der Strukturvergleichsprogramme konnte eine Vorhersagequalität von 4.8 zu 3.2 erreicht werden, das heißt 60% der Targets wurden richtig vorhergesagt. Bei den anderen Kriterien lag die Qualität zum Teil sogar erheblich höher (Klasse: 7.3 zu 0.7, Alignments: 5.3 zu 2.7).
Damit waren unsere Einreichungen sehr erfolgreich, das heißt unsere Gruppe lag in der Spitzengruppe von etwa 7 Arbeitsgruppen der beteiligten 35 Gruppen. Diese Einschätzung wird auch durch die mittlerweile veröffentlichten Ergebnisberichte der Evaluatoren [205, 217] der *Threading*–Kategorie des Wettbewerbs bestätigt.
Von diesen etwa sieben Gruppen war jede in der Lage, die meisten (das heißt mehr als drei Viertel) der möglichen ähnlichen Faltungen in der Datenbank der strukturaufgeklärten Proteine zu identifizieren und in vielen Fällen auch gute Alignments (mit kleiner *RMS*–Abweichung des Modells zur jetzt aufgeklärten Struktur) zu berechnen. Dies stellt einen entscheidenden Fortschritt innerhalb der letzten beiden Jahre seit `CASP I` dar. Ermutigend ist auch, daß doch recht unterschiedliche Verfahren heutzutage in der Lage sind, für sehr schwierige Probleme übereinstimmende, beziehungsweise gleich gute Lösungen zu produzieren. Allerdings wurde auch klar, daß viele der publizierten Methoden entscheidende Schwächen im praktischen Einsatz für relevante Datensätze aufweisen [118, 168, 194].
Leider haben alle Methoden noch ihre Schwächen. Keine der teilnehmenden Gruppen konnte alle möglichen Lösungen in der zur Verfügung stehenden Zeit bestimmen. Besondere Probleme bereiten die sogenannten „*false positives*“, das heißt zuversichtlichen Vorhersagen für Sequenzen, die sich als völlig falsch herausstellten, weil es überhaupt keine ähnlichen Strukturen in der Datenbank gab. Dieser

Target-Id.	# Reste	Beschreibung	Modell	Prob.	VAST	DALI	SSAP	NONE	Class	Fold	Topology	Structure	Alignment
t2	514	Threonine-Deaminase, E.coli	1PII-	0.80	0.80	0.80	-	0.80	0.80				
			1MIOD	0.20	0.20	0.20	-	0.20	0.20				
t4	84	Polyribonucleotide Nucleotidyltransferase, S1-motif,E.coli	1CSP-	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
			1SNC-	0.20	0.20			0.20	0.20	0.20	0.20	0.20	0.20
			1PGX-	0.40				0.40	0.40		0.40		
t14	252	3-Dehydroquinase, Salmonella-typhi	1TMHD	0.80	0.80		0.80	0.80	0.80	0.80	0.80	0.80	0.80
			1DBP-	0.20	0.20	0.20		0.20	0.20	0.20	0.20	0.20	
t20	320	Ferrochelatase, B.subtilis	5TIMB	0.40				0.40	0.40		0.40		
			NONE-	0.60									
t26	79	ArgR,N-terminal-domain, E.coli	1LEA-	0.90	X	X	X	0.90	0.90	0.90	0.90	0.90	0.90
			1LLIA	0.10	X	X	X	0.10	0.10	0.10	0.10	0.10	0.10
t31	242	Exfoliative-toxin-A, S.aureus	1TRY-	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
			NONE-	0.10									
t38	152	CBDN1,Cellulomonas-fimi	1EXG-	1.00		1.00			1.00	1.00	1.00	1.00	1.00
t42	78	NK-lysin,Pig	1CB1-	0.70				0.70	0.70	0.70	0.70		0.70
			NONE-	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
Ok	**GMD**	**PROTAL**	Total (8)	8	4.8	4.8	3.4	6.3	7.3	5.5	6.3	4.8	5.3
Wrong	**GMD**	**PROTAL**	Total (8)	8	3.2	3.2	3.6	1.7	0.7	2.5	1.7	3.2	2.7

Tabelle 8.12: CASP II: Erfolg der Vorhersagen für die 8 Strukturen mit ähnlichen Faltungen in der PDB bei Anwendung verschiedener Evaluierungsmethoden: Drei Spalten entsprechen den bei der Ergebnisbewertung eingesetzten Strukturvergleichsmethoden, fünf weitere Spalten der jeweiligen Übereinstimmung des vorhergesagten Modells mit der Struktur hinsichtlich Klasse, Faltung, Topologie, Struktur und Alignment. Eine schwarze oder farbige Box zeigt eine korrekte Vorhersage bezüglich des betrachteten Kriteriums an. Die Zahl in der Box ist die bei der Vorhersage geschätzte Wahrscheinlichkeit für die Korrektheit des Modells. Die Summen in den letzten beiden Zeilen geben die Gesamtwahrscheinlichkeit für die richtig bzw. falsch vorhergesagten Strukturen an [200].

Fall ist für *Threading*–Methoden besonders schwierig, da anhand der mit mehr oder weniger Zuverlässigkeit vorhergesagten Modelle entschieden werden muß, daß sie alle nicht zutreffend sind. Aber auch in diesem Bereich stellten sich unsere Vorhersagen als sehr erfolgreich heraus: Für die Hälfte dieser Targets (ein Spitzenwert im Rahmen des Wettbewerbs) konnte mit hoher Wahrscheinlichkeit das Vorliegen einer neuen Faltungsklasse vorhergesagt werden. Es konnte eine Genauigkeit von knapp unter 50% erreicht werden. Tabelle 8.13 faßt die Analyse unserer Vorhersagen für die sechs Proteine zusammen, zu denen bislang keine ähnliche Faltung in der Datenbank existierte.

Target-Id.	# Reste	Beschreibung	Modell	Prob	VAST	DALI	SSAP	NONE
t5	268	Gamma-Fibrinogen C-terminus,Human	1GGTA 2HWC1	0.60 0.40				
t10	456	Bactericial/Permeablility increasing-protein	1MIOD NONE-	0.20 0.80	0.80	0.80	0.80	0.80
t12	107	Proregion,Procaricain, Carica-papaya	1GCB- NONE-	0.30 0.70	0.70	0.70	0.70	0.70
t16	312	Peridinin-Chlorophyll Protein, Amphidinium-carterae	7ABP- 4ENL- NONE-	0.60 0.30 0.10	0.10	0.10	0.10	0.10
t22	591	L-Fucose-Isomerase, E.coli	1ANHB 1HLDB 1MAMH NONE-	0.20 0.20 0.20 0.40	0.20 0.40	0.20 0.40	0.40	0.20 0.20 0.20 0.40
t32	98	beta-cryptogein,fungus phytophthora-cryptogea	1LPT- NONE-	0.40 0.60	0.60	0.60	0.60	0.60
Ok	**GMD**	**PROTAL**	Total (6)	6	2.8	2.8	2.6	3.2
Wrong	**GMD**	**PROTAL**	Total (6)	6	3.2	3.2	3.4	2.8

Tabelle 8.13: CASP II: Übersicht über die Vorhersagen, bei denen die experimentell bestimmte Struktur sich als völlig neue Faltung herausgestellt hat (Einträge vergleiche Tabelle 8.12) [200].

Während der Vorhersage und auch retrospektiv bei der ersten Auswertung der Ergebnisse (insbesondere der fehlerhaften Vorhersagen) ergaben sich einige Ansätze für zukünftige Verbesserungen. Zum Beispiel hat sich die Einbeziehung multipler Alignments als sehr hilfreich erwiesen. Aus diesem Grunde ist es auch im Anschluß an diese Arbeit geplant, multiple Alignmentinformation in die RDP–Methode miteinzubeziehen, beziehungsweise die Methode zu einem neuartigen *multiplen Threadingverfahren* weiterzuentwickeln.

Die Fokussierung auf kompakte Alignments (das heißt Alignments mit wenigen Insertionen und Deletionen) hat sich zwar zur Faltungserkennung bewährt, führt jedoch in vielen Fällen zu schlechten Sequenzstrukturalignments mit hohen *RMS*–Abweichungen zwischen dem vorhergesagten Modell und der experimentell aufgeklärten Struktur. Ein Grund dafür liegt zum Beispiel darin, daß bei Verwendung hoher Kosten für Insertionen und Deletionen alle Schleifenbereiche aliniert werden, die bekanntlich auch zwischen homologen Strukturen sehr unterschiedlich sein können.

Da die RDP–Methode weitestgehend ohne explizite Gapkosten auskommt, enthalten die mit dieser Methode berechneten Sequenzstrukturalignments in der Regel mehr Insertionen und Deletionen. Diese bestehen jedoch in den meisten

Fällen aus sogenannten Wechselgaps, wo sowohl Bereiche aus der Sequenz als auch der Struktur unaliniert bleiben, da sie nicht zuverlässig einander zugeordnet werden konnten. So beschränkt die Methode automatisch das Alignment auf die strukturrichtig alinierbaren Bereiche von Sequenz und Struktur. Unterschiedliche Schleifenbereiche sind nicht Bestandteil des Alignments. Die in den folgenden Abschnitten durchgeführte Analyse dessen, welche Fortschritte durch die Anwendung der RDP–Methode auf die Beispiele des Wettbewerbs schon mit den Standardparametern der Methode erzielt werden können, zeigt dies deutlich. Durch Anwendung der RDP–Methode konnte so insbesondere die Qualität der Sequenzstrukturalignments in Bezug auf die bei der Bewertung der Vorhersagen angelegten Bewertungskriterien wesentlich verbessert werden.
Während der Vorhersagen für den Wettbewerb wurde dem Einfügen von Insertionen und Deletionen in der RDP–Methode nahezu vollkommen freier Lauf gelassen. Mit sehr niedrigen Gapkosten tendierte die RDP–Methode bei den verwendeten Kostenfunktionen manchmal dazu, allein die bezüglich der Kostenfunktion gut zueinander passenden Bereiche zu alinieren, ohne dabei zu berücksichtigen, inwiefern das Alignment in ein sinnvolles Strukturmodell mit geschlossener oder zumindest schließbarer Rückgratkette umsetzbar ist. Diese Beobachtung hat dazu geführt, daß die RDP–Methode in der Zwischenzeit durch Funktionen ergänzt wurde, die sowohl beim rekursiven Abstieg als auch beim Zusammensetzen der Teillösungen testen, ob eine Teillösung zu einem sinnvollen Strukturmodell erweitert werden kann (siehe Abschnitte 7.3.3 und 7.6.2). Zur Verbesserung der Sequenzstrukturalignments ist außerdem geplant, RDP um Modellierungsansätze von Schleifen, oder zumindest um zusätzliche Kriterien über ihre Modellierbarkeit, zu erweitern (siehe Abschnitt 9.1).
Im folgenden werden drei Vorhersagen im Detail diskutiert. Besondere Betonung soll dabei auf dem Anteil der RDP–Methode an dem erfolgreichen Abschneiden beim `CASP II`–Wettbewerb und auf die Verbesserungen gelegt werden, die durch die weiterentwickelte RDP–Methode erreicht und im nachfolgenden Wettbewerb erwartet werden können. Eine Diskussion aller Vorhersagen würde den Rahmen dieser Arbeit sprengen. Für eine Gesamteinordnung der gemachten Vorhersagen sei nochmals auf den Ergebnisberichte der Evaluatoren [205, 217] verwiesen.

8.5.1 Target t4: Polyribonukleotide Nukleotidyltransferase

Nach Aufklärung der Struktur der Nukleotidyltransferase [51] wurde als ähnlichste Struktur in der Strukturdatenbank die Struktur des *major cold shock* Proteins aus dem Organismus *Escherichia coli* [311] identifiziert. Aufgrund dieser strukturellen Ähnlichkeit wird vermutet, daß beide Proteine aus einem Nukleotide bindenden Protein als gemeinsamem Vorfahren hervorgegangen sind [51].
In den Suchläufen für `CASP II` war das Protein `1csp` für unterschiedliche Parametersätze und Paarinteraktionspotentiale in den Ranglisten auf den Positionen 4, 8, 9 und 16 von den 1334 Proteinen der verwendeten Repräsentativmen-

ge. Auf den Plätzen davor befanden sich jeweils größere Proteine, bei denen in den zugehörigen Alignments die Sequenz der Nukleotidyltransferase auf Strukturelemente abgebildet wurden, die ebenfalls als *OB*–Faltungen angesehen werden können. Aufgrund des kompakten Alignments, der relativ hohen Sequenzidentität (von 23.8%) und der Übereinstimmung mit den Ergebnissen der Erkennungsexperimente mit `123D` wurde `1csp` als mögliche Faltung der Nukleotidyltransferase identifiziert und eingereicht.
Mit der weiterentwickelten RDP–Methode und den in Abschnitt 7.7 ermittelten Parametern, wird `1csp` bereits auf Platz 3 der Rangliste gefunden. Direkt vor `1csp` in der Rangliste steht ein weiteres *major cold shock* Protein mit *OB*–Faltung (`1mjc`), das ebenfalls ein sehr gutes Modell für die Nukleotidyltransferase darstellt. Daß zwei Proteine gleicher Faltung am Anfang der Rangliste gefunden werden, würde die Signifikanz einer Blindvorhersage stark erhöhen, da es sich um die Rangliste bezüglich einer Repräsentativmenge handelt, in der nur wenige, untereinander nicht oder nur gering homologe *OB*–Faltungen enthalten sind.
Die wesentliche Verbesserung, die durch die Weiterentwicklungen der RDP–Methode erreicht wurde, ist die Qualität der berechneten Sequenzstrukturalignments. Abbildung 8.19 zeigt das Alignment, wie es für den Wettbewerb eingereicht wurde, und das Alignment, wie es von RDP mit den allgemein verwendeten Parametern heute berechnet wird. Die Entscheidung, das obere Alignment einzureichen, ist aufgrund seiner Kompaktheit und der für Faltungserkennungsexperimente sehr hohen Sequenzidentität von 24% gefallen. Die *RMS*-Abweichung von 5.2Å zeigt, daß in diesem Falle die Kompaktheit ein falsches Entscheidungskriterium war – zumindest wenn man die bei der Ergebnisbewertung angewendeten Kriterien anlegt. In dem RDP–Alignment werden zwei Wechselgaps so eingefügt, daß nur die wirklich passenden Bereiche einander zugeordnet werden, wobei neun Reste weniger aliniert werden. Da sich diese neun Reste in Wechselgaps befinden, ist gewährleistet, daß die Lücken in dem aus dem Alignment resultierenden Modell in einer anschließenden Modellierung problemlos geschlossen werden können. Das RDP–Alignment erreicht eine *RMS*–Abweichung von 2.9 Å bei einer sogar noch höheren Sequenzidentität von fast 27%.

	Identität [%]	# Reste	*RMS*–Abweichung [Å]
Sequenzalignment	24.5	67	7.3
`SARF`	21.0	64	2.3
`SARF-2`	12.0	48	2.6
Einreichung	23.8	67	5.2
RDP	26.9	58	2.9

Tabelle 8.14: Superposition der Struktur von `1csp` mit der experimentell ermittelten Struktur der Nukleotidyltransferase gemäß des RDP–Alignments

Tabelle 8.14 zeigt, daß diese so gute Werte selbst von Strukturalignmentmetho-

```
Submission for CASP 2:
1csp      00:-------MLEGKVKWFNSEKGFGFIEVEGQDDVFVHFSA--IQGEGFKTLEEGQAVSFEI
t04s      00:AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKV
score     00:       |  GKV || |  |F  I| | || V |   A    |   | L| GQ|V|  |
SEC '1csp'00:        eeeeeeeee    eeeee      eeeehhh  e            eeeeee
SEC 't04s'00:       eeeeeeeeee  eeeee                      hhhh     eeeeee

1csp      60:VEGNRGPQAANVTKEA--------
t04s      60:LEVDRQGRIRLSIKEATEQSQPAA
score     60:|E |R  |    |KEA
SEC '1csp'60:eee  eeeeeeeee
SEC 't04s'60:ee       eeee

Alignment                      value  Alignment    Prof-1    Prof-2    Mapped
Alignment length          =       84                 67        84        67
Alignment ids             =       16     19.05 %   23.88 %   19.05 %   23.88 %
Alignment homs            =       36     42.86 %   53.73 %   42.86 %   53.73 %
SecStruct Total (Gap = C) =       61     72.62 %             72.62 %

Alignment RMS             =     5.24

RDP sequence structure alignment:
1csp      _000:.......MLEGKVKWFNSEKGF--GFIEVEGQDDVFVHFSAIQGEGFKT---------LE
t04s      _000:AEIEVGRVYTGKVTRIVD----FGAFVAIGGGKEGLVHISQIAD-----KRVEKVTDYLQ
score     _000:.......   GKV            F    G     VH S I                L
SEC '1csp'_000:        eeeeeeeee    e  eeee      eeeehhhe
SEC 't04s'_000:       eeeeeeeeee      eeeee                          hhhh

1csp      _060:EGQAVSFEIVEGNRGPQAANVTKEA--------
t04s      _060:MGQEVPVKVLEVDRQGRIRLSIKEATEQSQPAA
score     _060: GQ V     E  R        KEA
SEC '1csp'_060:   eeeeeeeee  eeeeeeeee
SEC 't04s'_060:   eeeeeeee       eeee

target   PDB    score (energy [pos.] native   SEQ   CCP  del   ins)|len1|len2|#ali|id% |rms
t04s  ->1csp: -147.00 (  -9.7 [-0.2]  -15.0  99.0 178.0 -2.6 -20.6)|  84|  67|  58|26.9|2.9
                       *6.0                *-1.0 *-0.1 *0.2 *-0.2
```

Abbildung 8.19: Vergleich des bei CASP II von der PROTAL–Gruppe eingereichten (oben) mit dem von der RDP–Methode mit Standardparametern berechneten Sequenzstrukturalignment (unten).

den, wie SARF und SARF2, nicht immer erreicht werden können. So liefert zum Beispiel nur die alte Version von SARF ein besseres Strukturalignment.
Abbildung 8.20 zeigt die Superposition von 1csp mit der zwischenzeitlich aufgeklärten Proteinstruktur der Nukleotidyltransferase gemäß des mittels RDP berechneten Alignments. Die Superposition zeigt, daß die Alignmentqualität bei diesem Beispiel sicher groß genug ist, um auf seiner Basis den Versuch zu unternehmen, mit Hilfe der Methoden der vergleichenden Modellierung (siehe Abschnitt 4.1) ein alle Atome enthaltendes Strukturmodell zu erzeugen.

8.5.2 Target t14: 3-Dehydroquinase

Die Struktur der 3-Dehydroquinase aus *Salmonella typhi* wurde ebenfalls experimentell aufgeklärt, jedoch bisher nicht in der Proteinstrukturdatenbank veröffentlicht. Den Teilnehmern des Wettbewerbs stehen die Daten jedoch zur Verfügung, so daß die folgenden Untersuchungen möglich waren.
Abbildung 8.21 zeigt, daß bereits mit der ersten für die Vorhersagen im Wettbewerb eingesetzten Version der RDP–Methode die sogenannte *TIM–Barrel–*

Abbildung 8.20: Superposition der Struktur von `1csp` mit der experimentell ermittelten Struktur der Nukleotidyltransferase gemäß des RDP–Alignments

Faltung der Dehydroquinase eindeutig identifiziert werden konnte.

```
rank target   PDB     score [ sd ]( energy[pos.] native  hydro nathyd)|len1|len2| id% |  C   A   T  H         | SCOP

  0:t0014 ->1tpb2:-10560.08 [-1.8](  -61.3[-0.3]  -53.0  -15.1  -20.2)| 252| 247| 11.7|  3  20  40  TIM Barrel | BETA/ALPHA (TIM)-BARREL
  1:t0014 ->1tpvB: -2997.11 [-1.7](  -55.3[-0.3]  -51.3  -15.6  -22.1)| 252| 247| 10.9|  3  20  40  TIM Barrel | BETA/ALPHA (TIM)-BARREL
  2:t0014 ->1tpc2:  -729.42 [-1.7](  -56.9[-0.3]  -48.4  -14.7  -20.2)| 252| 247| 10.9|  3  20  40  TIM Barrel | BETA/ALPHA (TIM)-BARREL
  3:t0014 ->1tph2:  -640.32 [-1.7](  -59.1[-0.3]  -49.9  -15.3  -20.3)| 252| 247| 11.3|  3  20  40  TIM Barrel | BETA/ALPHA (TIM)-BARREL
  4:t0014 ->1tpuB:   617.19 [-1.7](  -54.7[-0.3]  -47.6  -13.9  -20.9)| 252| 247| 10.5|  3  20  40  TIM Barrel | BETA/ALPHA (TIM)-BARREL
  5:t0014 ->1nfp :  2760.15 [-1.7](  -85.8[-0.5] -103.2  -13.4  -12.8)| 252| 228| 13.6|  3  20  20  Flavoprot. | beta/alpha (TIM)-barrel
```

Abbildung 8.21: RDP–Rangliste für das Target t14: Gezeigt werden die ersten sechs Positionen der für die Blindvorhersage verwendeten Rangliste.

Die Tabelle 8.15 vergleicht den von der `PROTAL`–Gruppe in Abstimmung eingereichten Strukturvorschlag, das beste eingereichte Strukturmodell und das Strukturmodell, wie es gegenwärtig von der RDP–Methode mit Standardparametern berechnet würde.
Das eingereichte Strukturmodell basiert auf einem von `123D` mit der *TIM–barrel*–Struktur `1tmhD` berechneten Sequenzstrukturalignment und ordnet 80% der Reste der Dehydroquinase Strukturpositionen von `1tmhD` zu. Die beste Einreichung von allen Teilnehmern war mit einer *RMS*–Abweichung von 5.9Å ein Modell auf Basis der Struktur der Isomerase `1pii`. Die aktuelle Version der RDP–Methode berechnet mit Standardparametern dagegen ein Modell auf Basis der Lyasestruktur

	Modell	*RMS*–Abweichung [Å]	# alinierte Pos. [%]
`Protal` Einreichung	`1tmhD`	11.1	80.5
Beste Einreichung	`1pii`	5.9	89.7
RDP–Modell	`1nal4`	6.1	71.0

Tabelle 8.15: Vergleich verschiedener Strukturvorschläge für die 3-Dehydroquinase (Target t14)

`1nal4` mit einer *RMS*–Abweichung von 6.1Å zur Kristallstruktur. Damit wäre die vollkommen automatisch erzeugte RDP–Vorhersage nur unwesentlich schlechter als die manuell bearbeitete beste Vorhersage im Wettbewerb und würde damit heute einen Spitzenplatz belegen.
Abbildung 8.22 zeigt die Superposition der Dehydroquinasestruktur mit der Lyasestruktur `1nal4` gemäß des RDP–Sequenzstrukturalignments. Die Abbildung macht deutlich, daß durch das RDP–Modell die wesentlichen Struktureigenschaften, so zum Beispiel das zentrale *β–Barrel* und die in beiden Strukturen vorhandenen Helices, korrekt modelliert werden.

8.5.3 Target t31: Exfoliatives Toxin

Für das Target t31 war sowohl aus den `123D`–Ranglisten als auch aus den RDP–Ranglisten eindeutig zu erkennen, daß es sich bei der Faltung des exfoliativen Toxins aus *S. aureus* um eine den Serinproteasen ähnliche Faltung handelt. Daher wurde mit sehr hoher Wahrscheinlichkeit (0.9) ein Modell auf Basis der Struktur der Hydrolase `1try` vorhergesagt. Diese Vorhersage wurde durch die experimentell aufgeklärte Struktur bestätigt.
Zur Berechnung des eingereichten Alignments wurden, wie bereits diskutiert, relativ hohe Kosten für Insertionen und Deletionen verwendet. Das Ergebnis war ein relativ kompaktes Alignment, das 189 der 242 Aminosäurereste des Toxins Strukturpositionen der Hydrolase `1try` zuordnet. Der Nachteil dieses Alignments mit einer Sequenzidentität von 21% ist die relativ hohe *RMS*–Abweichung zur zwischenzeitlich aufgeklärten Struktur des Toxins von 7.4Å. Im Unterschied dazu wies das beste für `CASP II` eingereichte Modell auf Basis der Hydrolase `1ppfE` eine *RMS*–Abweichung von nur 4.2Å bei 190 zugeordneten Positionen auf.
Abbildung 8.23 zeigt das mit der gegenwärtigen Version der RDP–Methode und den Standardparametern berechnete Sequenzstrukturalignment des exfoliativen Toxins (Target t31) mit der Hydrolase `1try`. In diesem Alignment werden 150 Reste des Toxins Strukturpositionen der Hydrolase mit einer *RMS*–Abweichung von 3.9Å zugeordnet. Das RDP–Sequenzstrukturalignment zeigt in den Bereichen um die für die enzymatische Aktivität einer Serinprotease verantwortlichen Reste der katalytischen Triade eine weitaus höhere Konservierung der Sequenz, als dies für die restlichen Bereiche der Fall ist. Das Histidin, das Aspartat und das Serin der katalytischen Triade werden durch das Alignment konserviert und sind in

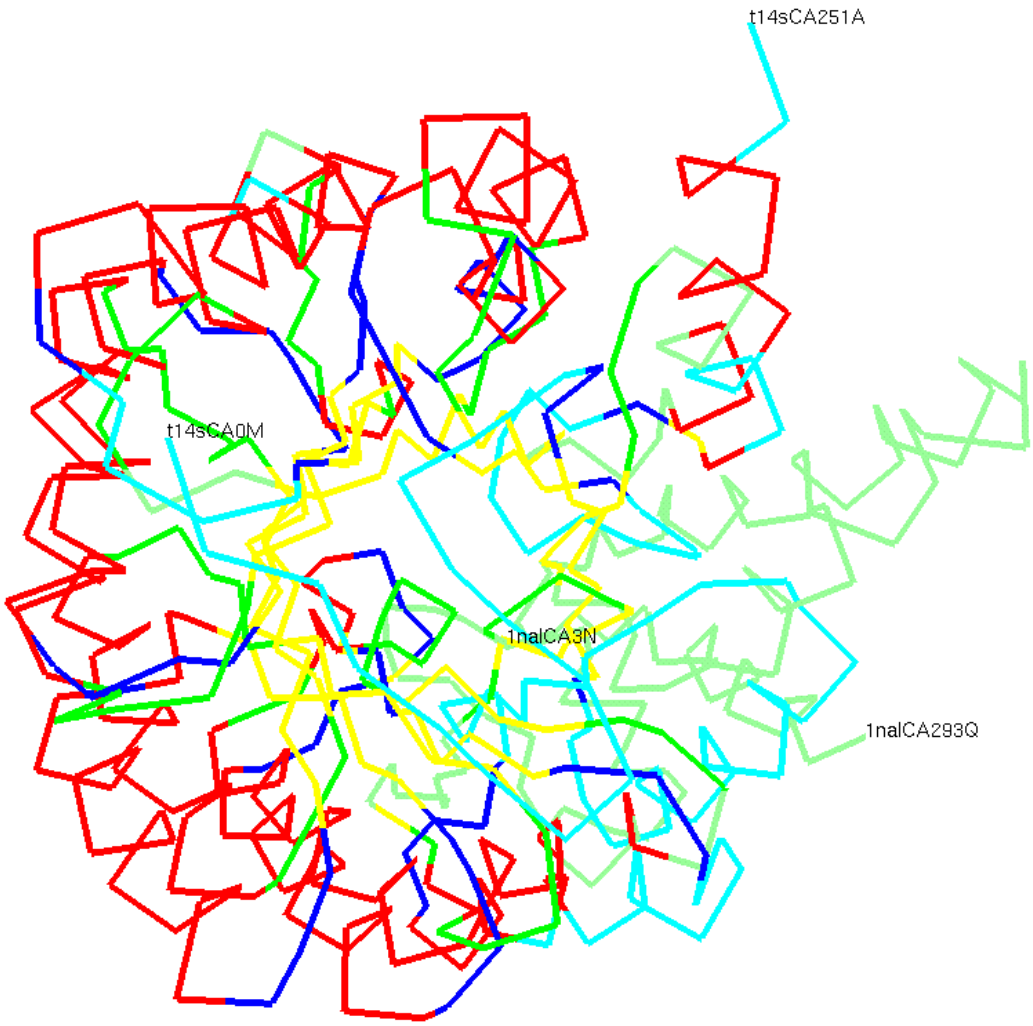


Abbildung 8.22: Superposition der Dehydroquinase mit `1nal4` gemäß des RDP–Sequenzstrukturalignments

Abbildung 8.23 rot hervorgehoben.

Die Konserviertheit der katalytischen Triade legt die Vermutung nahe, daß auch das Toxin eine katalytische Aktivität hat. Damit kann in diesem Fall über die vorhergesagte strukturelle Ähnlichkeit auf eine mögliche Funktion geschlossen werden.

Die Superposition der experimentell ermittelten Toxinstruktur mit der Hydrolasestruktur `1try` gemäß des RDP–Alignments in Abbildung 8.24 zeigt, daß die Konformation der katalytischen Triade auch strukturell konserviert ist. Die Reste der jeweiligen katalytischen Triaden sind durch Beschriftung hervorgehoben. Wie in so vielen Fällen ist für dieses Beispiel das Strukturmodell gerade im akti-

```
1try      _000:----------------IVGGTSASAG--------------------DFPFIVSISRNGGPWCGGSLLNANTVLTAAHCVSGYAQSG--------FQIRA
t31s      _000:EVSAEEIKKHEEKWNKYYGV------NAFNLPKELFSKVDEKDRQKYPYNTIGNVFVKGQTSATGVLIGKNTVLTNRHIAK------FANGDPSKVSFRP
score     _000:                ||G                            |  I  |   G      G L   NTVLT  H|                    R
SEC '1try'_000:                                                  eeeeee  eeeeeeeeee  eeee hhhh    hhh         eeeee
SEC 't31s'_000:   hhhhhhhhhhhhhh         hhh     eee    hhh   hhheeeeee   eeeeeee    eeee hhhhh      hh   hhheeeee

1try      _100:GSLSRTSGGIT--------------SSLSSVRVHPSYS----GNNNDLAILKLST-----SIPSGGNIGYARLAASGSDPVAGSSATVAGWGATSEGGSS
t31s      _100:SI---------NTDDNGNTETPYGEYEVKEIL------QEPFGAGVDLALIRL--KPDQNGVSLGDKISPAKIGTSNDLKDGDKLELIGY----------
score     _100:                              |          G   DLA  |L        |  G  I  A|   S           |
SEC '1try'_100:         ee              e eeeeee              eeeee                                 eeeeee
SEC 't31s'_100:                         eeeeeee               eeeee          hhh         hhh     eeeeee

1try      _200:T-----PVNLLK----VTVPIVSRATCRAQYGTSAITNQM-FCAGVSSGGKD--SCQGDSGGPIVDSSNTLIGAVSWG---NGCARPNY-----SGVYAS
t31s      _200:-PFDHKVNQM--HRSEIELTTLSR----------------GLRYY-------GFTVPGNSGSGIFNSNGELVGIHS--SKVSHL-----DREHQINYGVG
score     _200:        ||      |     SR                 ||            | G SG  I  S   L|G  S                      |
SEC '1try'_200:           ee    eeeeee hhhhhhhh        e eee           eeee    eeeeeee                    eeeee
SEC 't31s'_200:             eeeeeeee hhh                 eeee           hhh   eee    eeeeeee  eeee           eeeeeee

1try      _300:VGALRSFIDTYA----     Alignment                      value  Alignment    Prof-1    Prof-2    Mapped
t31s      _300:IGNYVKRIIN--EKNE     Alignment length           =     316                224       242       150
score     _300:|G     I |           Alignment ids              =      33     10.44 %   14.73 %   13.64 %   22.00 %
SEC '1try'_300:hhhhhhhhhhh          Alignment gaps             =      25      7.91 %   11.16 %   10.33 %   16.67 %
SEC 't31s'_300:  hhhhhhhh  hh       Alignment RMS              =    3.91 ( 150)
```

Abbildung 8.23: RDP–Alignment des exfoliativen Toxins gegen die Struktur der Hydrolase `1try`.

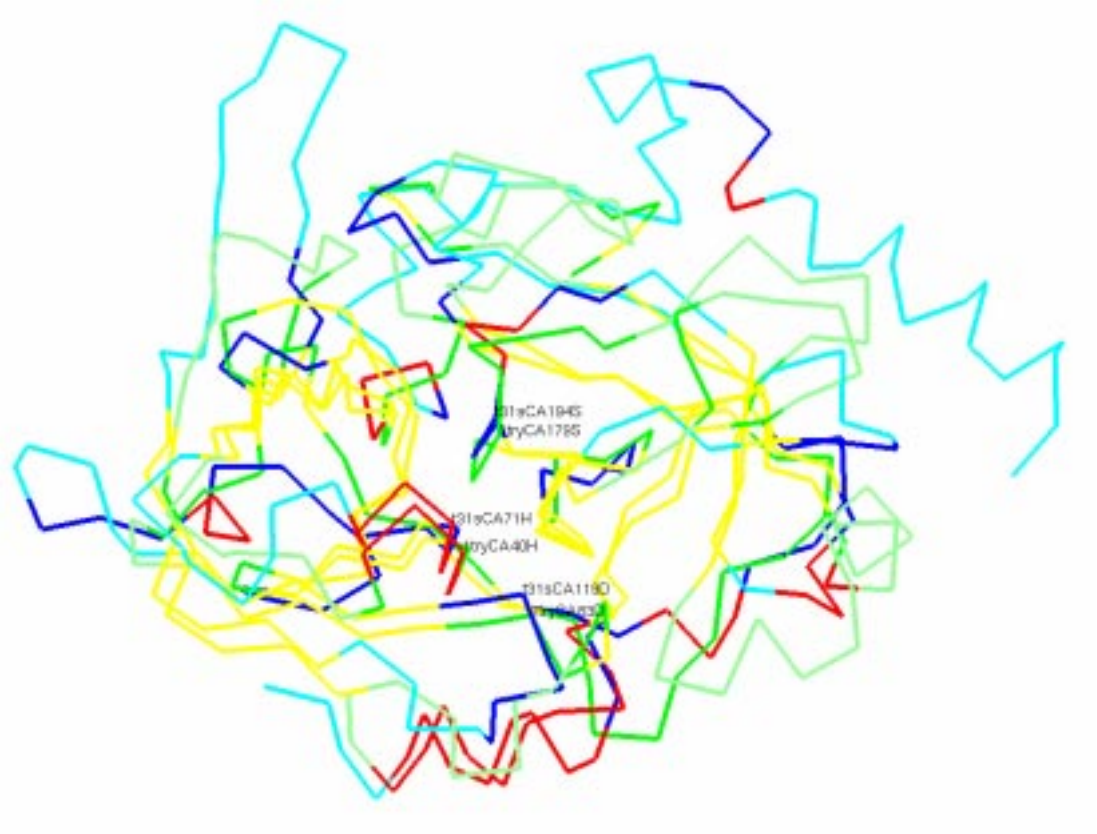


Abbildung 8.24: Superposition des Toxins mit `1try` gemäß des RDP–Sequenzstrukturalignments.

ven Zentrum des Proteins besser, als es die *RMS*–Abweichung über das gesamte Protein erwarten läßt. Das aktive Zentrum ist jedoch der Bereich des Proteins, der für auf die Strukturvorhersage aufbauende Untersuchungen in der Regel von wesentlichem Interesse ist.

8.6 Strukturvorhersage für die Thymidinkinase

Abbildung 8.25 zeigt die 376 Aminosäurereste lange Sequenz eines Proteins, an dessen Beispiel an der GMD im November 1995 die erste echte *Blindvorhersage* durchgeführt wurde [381]. Für die Vorhersage eines ersten strukturellen Modells

```
us_000:MASYPGHQHASAFDQAARSRGHSNRRTALRPRRQQEATEVRPEQKMPTLL
us_050:RVYIDGPHGMGKTTTTQLLVALGSRDDIVYVPEPMTYWRVLGASETIANI
us_100:YTTQHRLDQGEISAGDAAVVMTSAQITIGMPYAVTDAVLAPHIGGEAGSS
us_150:HAPPPALTLIFDRHPIAALLCYPAARYLMGSMTPQAVLAFVALIPPTLPG
us_200:TNIVLGALPEDRHIDRLAKRQRPGERLDLAMLAAIRRVYGLLANTVRYLQ
us_250:GGGSWREDWGQLSGTAVPPQGAEPQSNAGPRPHIGDTLFTLFRAPELLAP
us_300:NGDLYNVFAWALDVLAKRLRPMHVFILDYDQSPAGCRDALLQLTSGMIQT
us_350:HVTTPGSIPTICDLARTFAREMGEAN
```

Abbildung 8.25: Proteinsequenz (376 Aminosäuren) unbekannter Struktur.

für die gezeigte Sequenz wurden im wesentlichen die in Abschnitt 8.2 bereits beschriebenen Arbeitsschritte durchgeführt. Dazu muß angemerkt werden, daß die Vorhersage unter großem Zeitdruck stattfand, da die Veröffentlichung der von zwei unabhängigen Arbeitsgruppen experimentell bestimmten dreidimensionalen Struktur [45, 363] kurz bevorstand und so nur etwa 10 Tage für eine echte Blindvorhersage zur Verfügung standen. Mittels einer einfachen `BLAST`–Datenbanksuche (siehe Abbildung 8.26) konnte die Proteinsequenz als eine Thymidinkinase (`TK`) des *Herpes Simplex Virus I* identifiziert werden. Leider war zum Zeitpunkt der Untersuchung für keine virale Thymidinkinase die Struktur aufgeklärt, beziehungsweise nicht in der Strukturdatenbank abgelegt, so daß die bei der einfachen Datenbanksuche gefundenen Proteine alle aus Sequenzdatenbanken wie zum Beispiel `SwissProt` [19] und `PIR` [111] stammten und keine Hinweise auf die dreidimensionale Struktur zuließen.
Nachdem bekannt war, daß die Sequenz eine Thymidinkinase des *Herpes Simplex Virus I* ist, konnte das Protein über seine *EC*–Nummer mittels `KEGG` [174] als ein Enzym im Pyrimidin–Metabolismus identifiziert werden. Abbildung 8.27 zeigt den gesamten Metabolismus. Der Teilpfad, an dem die Thymidinkinase beteiligt ist, ist dick in dem metabolischen Netzwerk hervorgehoben.
Die Thymidinkinase phosphoryliert Thymidin zu Thymidin–5'–monophosphat (Thymidylat), wobei Adenosintriphosphat (ATP) zu Adenosindiphosphat (ADP) degradiert wird. Diese Phosphorylierung stellt den ersten Schritt im „Recycling" von Thymidin für die DNA–Synthese dar. Die ebenfalls in Abbildung 8.27 hervorgehobene DNA–Polymerase verwendet nach zwei weiteren Phosphorylierungsschritten Thymidin–5'–triphosphat zum Aufbau einer neuen DNA–Kette. Ohne dieses „Recycling" von Thymidin ist eine Vermehrung des *Herpes Simplex Virus* unmöglich. Die Thymidinkinase ist daher auch ein seit geraumer Zeit bekanntes Zielprotein für die Bekämpfung viraler Herpesinfektionen. Aciclovir ist der bekannteste Wirkstoff, der zur Bekämpfung des Infektes eingesetzt wird. Aciclovir wird wie Thymidin von der Thymidinkinase phosphoryliert und tritt nach

```
                                                                  High  Probability
Sequences producing High-scoring Segment Pairs:                  Score  P(N)      N

sp|P08333|KITH_HSV1E   THYMIDINE KINASE (EC 2.7.1.21). >p...  1962  2.6e-269  1
sp|P06479|KITH_HSV1S   THYMIDINE KINASE (EC 2.7.1.21). >p...  1959  6.9e-269  1
pir|A93715|KIBET       thymidine kinase (EC 2.7.1.21) - h...  1941  2.2e-266  1
sp|P17402|KITH_HSV1K   THYMIDINE KINASE (EC 2.7.1.21). >p...  1929  1.0e-264  1
sp|P03176|KITH_HSV11   THYMIDINE KINASE (EC 2.7.1.21). >p...  1927  2.0e-264  1
sp|P06478|KITH_HSV1C   THYMIDINE KINASE (EC 2.7.1.21). >p...  1926  2.7e-264  1
gp|J04327|HS1TKM_1     Herpes simplex virus type 1 (mutan...  1710  3.6e-234  1
gp|V00466|HEHS07_1     Herpes simplex virus gene coding f...  1265  3.6e-200  2
gp|M29941|HS2TK2A_1    Herpes simplex virus type 2 (strai...  1260  1.8e-199  2
gp|S63520|S63520_1     thymidine kinase [herpes simplex v...  1256  6.5e-199  2
gp|X03896|HEHSV1GH_1   Herpes simplex virus type 1 gene f...  1192  6.7e-162  1
sp|P04407|KITH_HSV23   THYMIDINE KINASE (EC 2.7.1.21). >p...   881  7.9e-119  1
gp|K02122|HSVTKY_1     thymidine kinase [Marmoset herpesv...   313  2.0e-99   4
sp|P22649|KITH_HSVBH   THYMIDINE KINASE (EC 2.7.1.21). >p...   552  3.7e-92   3
gp|S46714|S46714_1     thymidine kinase [herpes simplex v...   458  7.5e-88   2
sp|P27363|KITH_PRVN3   THYMIDINE KINASE (EC 2.7.1.21). >g...   292  8.9e-82   4
gp|J03366|HS1TKA_1     Herpes simplex virus type 1 thymid...   609  1.8e-80   1
sp|P13159|KITH_HSVF    THYMIDINE KINASE (EC 2.7.1.21). >p...   266  2.1e-73   4
gp|M29943|HS2TK4A_1    Herpes simplex virus type 2 (TK- s...   352  3.6e-73   2
sp|P36226|KITH_HSVB5   THYMIDINE KINASE (EC 2.7.1.21). >p...   257  5.2e-68   6
gp|X75765|HVATK_1      thymidine kinase [unidentified]         252  2.1e-63   3
sp|P24424|KITH_HSVBQ   THYMIDINE KINASE (EC 2.7.1.21). >p...   284  9.8e-59   5
gp|U25806|CHU25806_1   thymidine kinase [Cercopithecine h...   230  2.7e-53   3
sp|P14343|KITH_VZVG    THYMIDINE KINASE (EC 2.7.1.21). >p...   212  8.8e-51   4
sp|P14344|KITH_VZVW    THYMIDINE KINASE (EC 2.7.1.21). >p...   212  8.8e-51   4
sp|P14341|KITH_VZV4    THYMIDINE KINASE (EC 2.7.1.21). >p...   209  3.3e-50   4
sp|P14342|KITH_VZV7    THYMIDINE KINASE (EC 2.7.1.21). >p...   208  5.0e-50   4
sp|P09250|KITH_VZVD    THYMIDINE KINASE (EC 2.7.1.21). >p...   212  1.2e-49   4
sp|P09100|KITH_HSVEB   THYMIDINE KINASE (EC 2.7.1.21). >p...   300  4.4e-48   2
sp|P24425|KITH_HSVE4   THYMIDINE KINASE (EC 2.7.1.21). >p...   299  4.4e-48   2
pir|S25497|S25497      thymidine kinase (EC 2.7.1.21) (mu...   294  3.7e-47   2
gp|S70449|S70449_1     thymidine kinase [Unknown.]             364  1.0e-46   2
sp|P24096|KITH_HSVB6   THYMIDINE KINASE (EC 2.7.1.21). >p...   253  6.4e-46   4
sp|P04408|KITH_HSVMR   THYMIDINE KINASE (EC 2.7.1.21). >p...   181  2.1e-45   3
sp|P13157|KITH_HSVTF   THYMIDINE KINASE (EC 2.7.1.21). >p...   140  1.4e-33   4
sp|P25987|KITH_HSVTU   THYMIDINE KINASE (EC 2.7.1.21). >p...   140  7.7e-33   4
gp|A04086|A04086_4     Synthetic MDV genes TK, gH and fla...   127  3.4e-30   4
```

Abbildung 8.26: BLAST–Datenbanksuche mit der unbekannten Sequenz: Thymidinkinase des *Herpes Simplex Virus I.*

zwei weiteren Phosphorylierungen als acyclo–Guanosintriphosphat (GTP) an die Stelle des Thymidin–5'–triphosphates und hemmt bevorzugt die virale DNA–Polymerase [191] und führt damit zum Kettenbruch in der synthetisierten DNA. Der große Vorteil dieses Wirkstoffes ist, daß die Anknüpfung des ersten Phosphatrestes nur von der *viralen* Thymidinkinase erfolgen kann, so daß sich eine hohe Selektivität auf mit dem Virus befallene Zellen ergibt und so gesunde Zellen nicht geschädigt werden. Aufgrund dieser hohen Selektivität wird Aciclovir zusammen mit der über einen Vektor in die befallene Zelle gebrachten viralen Thymidinkinase in der Krebstherapie genutzt.

Dies ist neben der weiten Verbreitung des Herpesvirus in der Bevölkerung ein weiterer Grund, weshalb man an der Entwicklung eines noch besseren Wirkstoffs interessiert ist. Um aber einen besseren Wirkstoff zielgerichtet entwerfen zu können, bedürfte es unter anderem der Kenntnis der dreidimensionalen Struktur der TK. Aus diesem Grund war die TK über viele Jahre ein Ziel für die Anwendung experimenteller und theoretischer Strukturvorhersagemethoden [126].

Da die Datenbanksuche mit Standardverfahren nach bekannten verwandten Strukturen erfolglos war, wurden im nächsten Schritt Faltungserkennungsläufe mit der

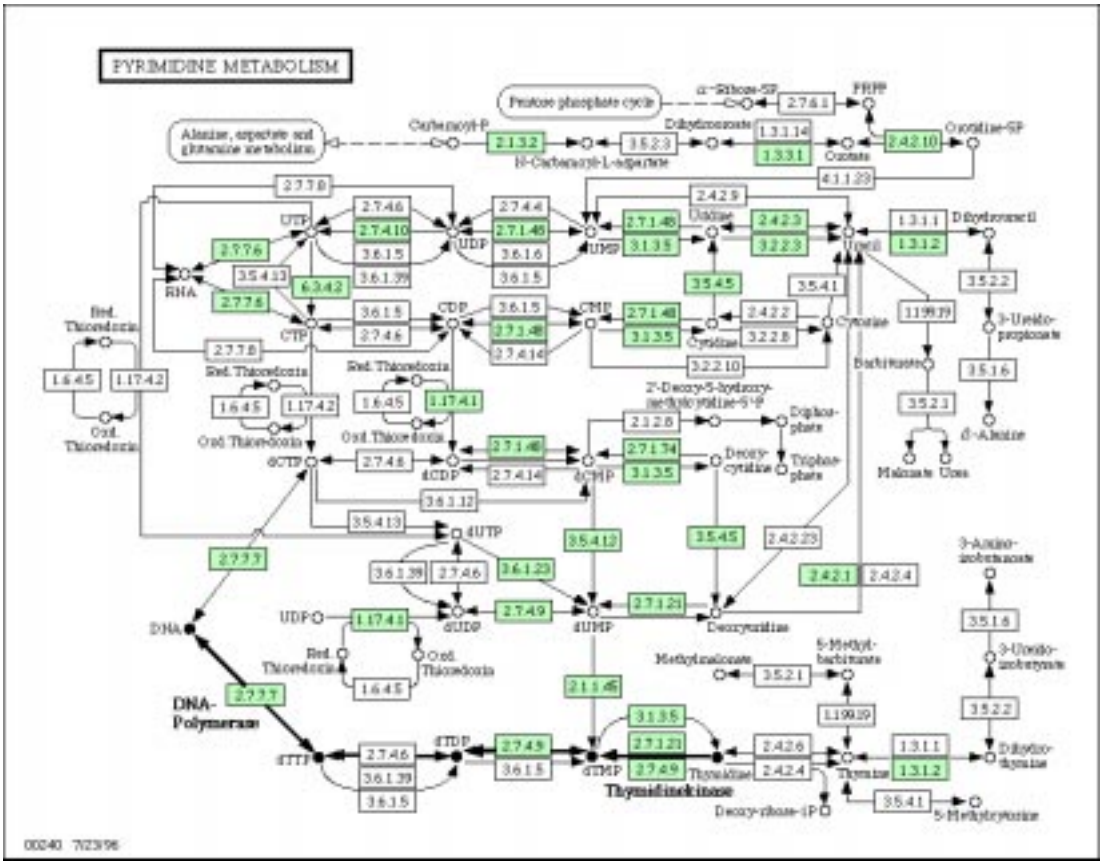


Abbildung 8.27: Einbindung der Thymidinkinase in den Pyrimidin–Metabolismus [174].

123D–Methode gegen die damalige Version der Strukturdatenbank PDB [29] mit verschiedenen Parametersätzen durchgeführt. Abbildung 8.28 zeigt für einen typischen Faltungserkennungslauf die Ranglistenpositionen der Proteine in der PDB nach CATH–Klassen [228] (siehe Abschnitt 3.3.5.1) sortiert. Eine genauere Analyse der in Abbildung 8.28 aufbereiteten Daten zeigte, daß auf den vorderen Positionen in der Rangliste im wesentlichen Proteine mit alternierenden helikalen und gestreckten Sekundärstrukturelementen und darunter vorrangig Proteine der CATH–Klasse „α/β *doubly wound*„ zu finden waren. Aufgrund dieser Ergebnisse erfolgte eine erste Eingrenzung des Suchraumes auf die Proteine dieser Klasse.

Die Richtigkeit dieser einschränkenden Annahme wurde zudem durch verschiedene Sekundärstrukturvorhersageprogramme [113, 288, 289] bestätigt. Abbildung 8.29 zeigt eine Konsensusvorhersage für die Sekundärstruktur der TK (mit SECSTR tk markierte Zeilen) im Vergleich mit der aus der zwischenzeitlich aufgeklärten Struktur ermittelten Sekundärstruktur (SECSTR tk_A markierte Zeilen).

Der Vergleich der Vorhersage mit der wirklichen Sekundärstruktur zeigt, wie zuverlässig Sekundärstrukturvorhersagen in manchen Fällen sind. Die Sekundärstrukturvorhersage läßt bereits deutlich die fünf Stränge erkennen, die – wie man

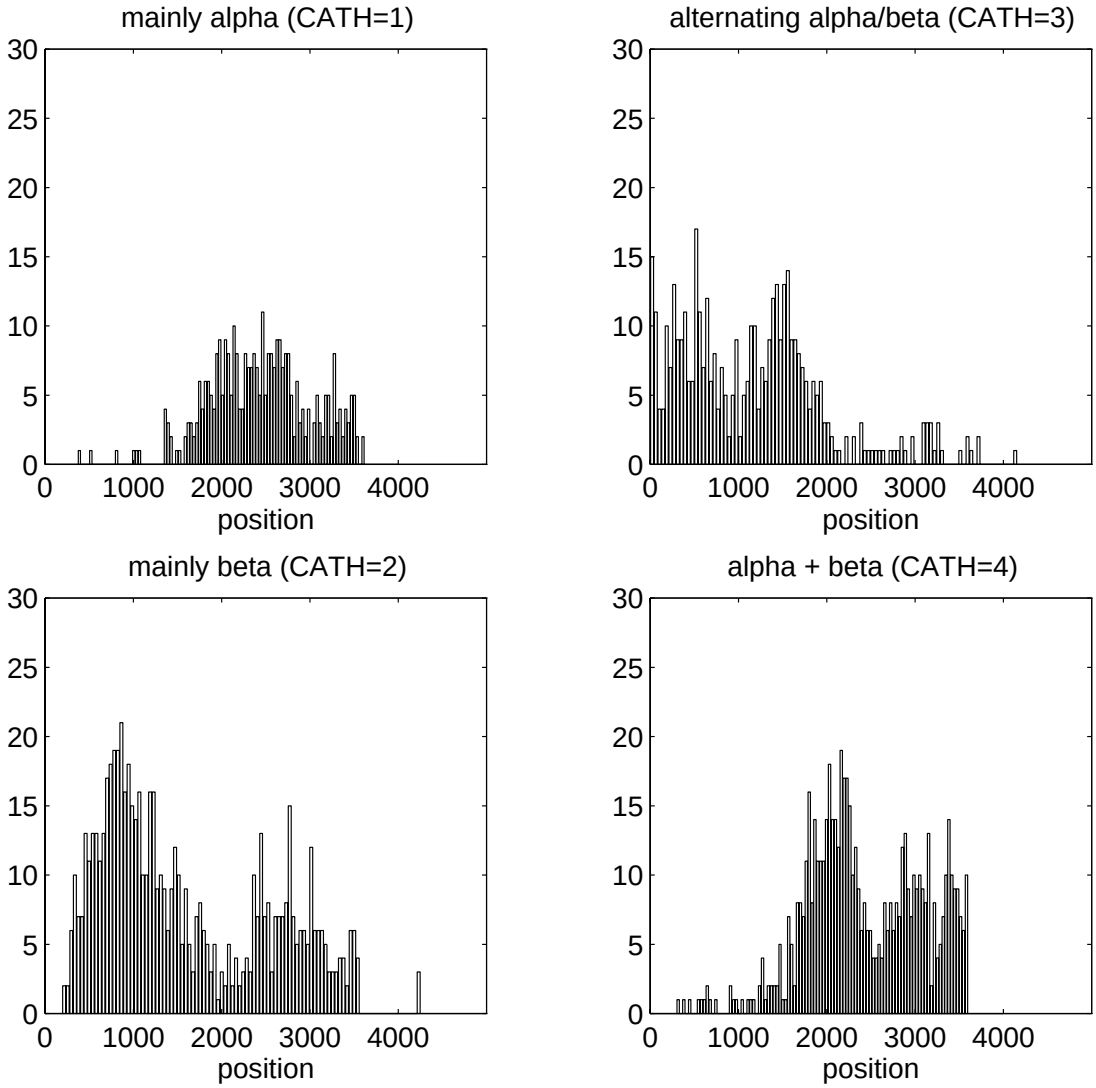


Abbildung 8.28: Faltungserkennung mit 123D: Positionen in der Rangliste getrennt nach den CATH–Klassen: hauptsächlich α, alternierenden α und *beta*, hauptsächlich β und $\alpha + \beta$ Proteine.

jetzt weiß – den zentralen Bereich der dreidimensionalen Struktur ausmachen.

Da die Thymidinkinase enzymatisch an der Phosphorylierung von Thymidin beteiligt ist und dabei sowohl mit dem Thymidin als auch mit ATP wechselwirken muß, wurde mit PROSITE [17] nach möglichen Bindestellen gesucht, die bereits auf der Sequenzebene erkennbar sind. Die Suche nach potentiellen Bindungsstellen in der TK–Sequenz ergab eine ATP–Bindestelle, die durch das Sequenzmuster `[AG]-x(4)-G-K-[STG]` charakterisiert ist. Dieses Sequenzmuster ist hoch konserviert und wird aufgrund seiner Funktion und Struktur auch *P–Loop* genannt. Inzwischen ist bekannt, daß ATP–bindende Proteine mit unterschiedlichen Faltungen ein gemeinsames strukturelles Motiv zur ATP–Bindung ausbilden [182].

```
tk           _000:MASYPGHQHASAFDQAARSRGHSNRRTALRPRRQQEATEVRPEQKMPTLLRVYIDGPHGMGKT
tk_A         _000:-------------------------------------------MPTLLRVYIDGPHGMGKT
SECSTR 'tk'  _000:CCCCCCCCCHHHHHHHHHCCCCCCCHHHCCCCCHHHHHHHCCH-CCCHEEEEEEECCCCCCCC
SECSTR 'tk_A'_000:                                                eeeeeee       hh

tk           _063:TTTQLLVALGSRDDIVYVPEPMTYWRVLGASETIANIYTTQHRLDQGEISAGDAAVVMTSAQI
tk_A         _063:TTTQLLVALGSRDDIVYVPEPMTYWRVLGASETIANIYTTQHRLDQGEISAGDAAVVMTSAQI
SECSTR 'tk'  _063:CHHHHHHH-CCCCCEEECCCCCC-EE--CCCHHHHHHH-C--CCCCCCCCH-HHHHHHHHHH-
SECSTR 'tk_A'_063:hhhhhhhh      eeee    hhhhh      hhhhhhhhhhhhh     hhhhhhhhhhhhh

tk           _126:TIGMPYAVTDAVLAPHIGGEAGSSHAPPPALTLIFDRHPIAALLCYPAARYLMGSMTPQAVLA
tk_A         _126:TMGMPYAVTDAVLAPHIGGEAG-----PPALTLIFDRHPIAALLCYPAARYLMGSMTPQAVLA
SECSTR 'tk'  _126:CCCCCCHHHHHHHHCCCCCCCCCCCCCCCEEEEE-CCCCHHHHHHCC-HHE-ECCCCHHHHHH
SECSTR 'tk_A'_126:hhhhhhhhhhhhhhhheeee          eeeeee  hhhhh hhhhhhhh     hhhhhh

tk           _189:FVALIPPTLPGTNIVLGALPEDRHIDRLAKRQRPGERLDLAMLAAIRRVYGLLANTVRYLQGG
tk_A         _189:FVALIPPTLPGTNIVLGALPEDRHIDRLAKRQRPGERLDLAMLAAIRRVYGLLANTVRYLQGG
SECSTR 'tk'  _189:HHH-CCCCCCCCCEEEECCCCHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHHHHHHHH-E-CC
SECSTR 'tk_A'_189:hh          eeeeee  hhhhhhhhhh          hhhhhhhhhhhhhhhhhhhhhh

tk           _252:GSWREDWGQLSGTAVPPQGAEPQSNAGPRPHIGDTLFTLFRAPELLAPNGDLYNVFAWALDVL
tk_A         _252:GSWREDWGQLSG---------------RPHIGDTLFTLFRAPELLAPNGDLYNVFAWALDVL
SECSTR 'tk'  _252:CCC---CCCCCCCCCCCCCCCCCCCCCCC-C-HHHHHCCCCCCCCCCCCCHHHHHHHHHH
SECSTR 'tk_A'_252:  hhhh hhhh                    hhh hhhhh             hhhhhhhhh

tk           _315:AKRLRPMHVFILDYDQSPAGCRDALLQLTSGMIQTHVTTPGSIPTICDLARTFAREMGEAN
tk_A         _315:AKRLRPMHVFILDYDQSPAGCRDALLQLTSGMIQTHVTTPGSIPTICDLARTFAREMGEAN
SECSTR 'tk'  _315:HHHHHCH-EEEECCCCCCCCHHHHHHHHHHH----CCCCCCCC--HHHHHHHHHHHHHHH
SECSTR 'tk_A'_315:hhhh   eeeeee    hhhhhhhhhhhhh    eee    hhhhhhhhhhhhhhh
```

Abbildung 8.29: Sekundärstrukturvorhersage für die TK: Vorhersage (SECSTR tk–Zeilen) *vs.* Struktur (SECSTR tk_A–Zeilen).

Da die ATP–Bindung grundlegend für die Funktion der TK ist, wurde der Suchraum weiter auf die Proteinstrukturen eingeschränkt, die zu der bereits bestimmten Faltungsklasse gehörten und zusätzlich eine ATP–Bindungstelle hatten. Über diese Filterkriterien konnte 1995 eine Menge von 28 Proteinstrukturen in der PDB identifiziert werden, die als plausible Modellvorlagen für die TK dienen konnten. In Faltungserkennungsläufen auf dieser eingeschränkten Menge konnten mit der RDP–Methode die Adenylatkinasen vom Schwein (3adk [85], 194 Aminosäurereste) und vom Rind (2ak3 [79]), 226 Aminosäurereste) als plausibelste Modellvorlagen identifiziert werden. Hierbei war von Vorteil, daß die RDP–Methode weitestgehend ohne die Bestrafung von Insertionen und Deletionen auskommt, da beide Modellvorlagen mit 226 beziehungsweise 194 Aminosäureresten wesentlich kürzer als die 376 Aminosäurereste lange TK sind und somit bei einem Alignment mindestens 150 Insertionen notwendig werden.

Abbildung 8.30 zeigt das RDP–Sequenzstrukturalignment der TK gegen die Struktur der Adenylatkinase 2ak3B. Die *P–Loop* der TK ist in diesem Alignment auf die *P–Loop* der Adenylatkinase abgebildet worden. Auch die vorhergesagten Sekundärstrukturelemente korrelieren in dem vorliegenden Alignment sehr gut mit denen der Modellvorlage. So konnten vier der fünf vorhergesagten Stränge auf Stränge der Struktur abgebildet werden. Daß nicht alle Helices abgebildet werden konnten, ist bei dem Längenunterschied nicht verwunderlich.

Die Hypothese, eine Adenylatkinase als Modellvorlage zu verwenden, wurde auch durch die Analyse der Zuverlässigkeit des Alignments in den Bereichen der Stränge mit den **ToPLign**–Methoden zur Zuverlässigkeitsanalyse [227] basierend auf den

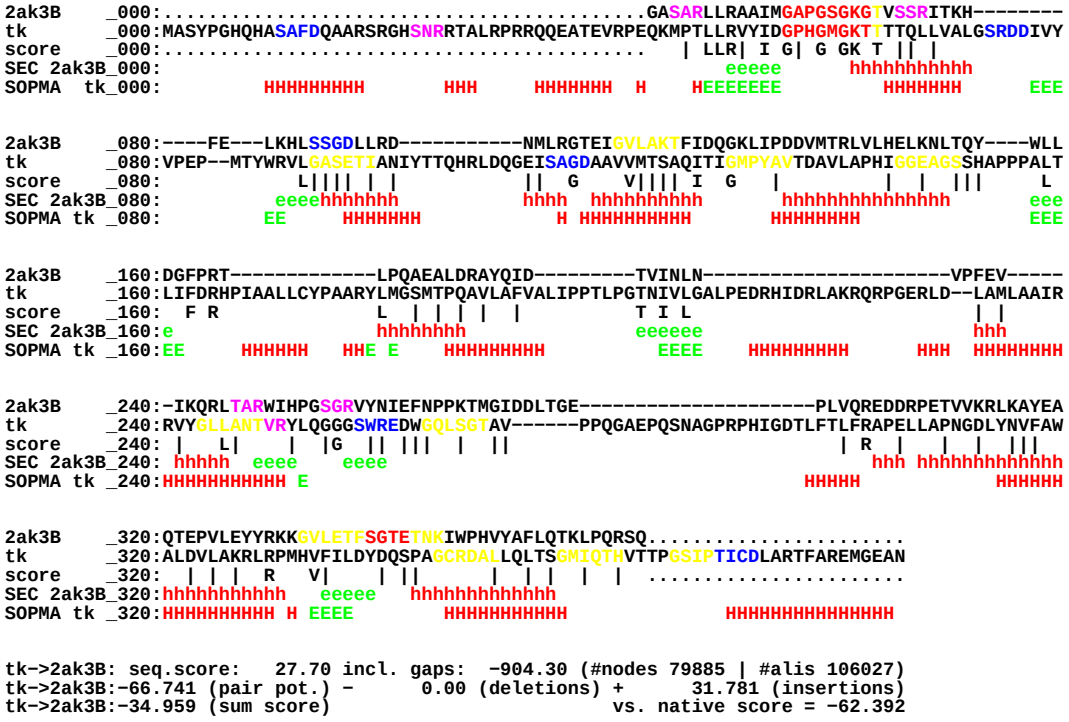


Abbildung 8.30: RDP–Sequenzstrukturalignment der TK gegen die Struktur der Adenylatkinase 2ak3B. Das Alignment ist annotiert mit der Sekundärstruktur von 2ak3B und der Sekundärstrukturvorhersage für die TK (*P–Loop*'s sind rot hervorgehoben.)

Pfadprofilmatrizen der 123D–Methode [5] untermauert [381]. Die Zuverlässigkeitsanalyse zeigte jedoch auch, daß in dem 123D–Alignment ein klarer Fehler im Bereich des letzten β–Stranges existierte, der ein noch besseres Modell verhinderte. Die anderen Proteine – keine Adenylatkinasen – die im oberen Bereich der RDP–Rangliste zu finden waren, konnten durch die zu dem Zeitpunkt noch manuell durchgeführte Analyse der zugehörigen Alignments als Modellvorlagen weitestgehend ausgeschlossen werden. Daher wurde auf der Grundlage des RDP–Alignments mit dem Programm MODELLER [305] und manuellen Eingriffen ein erstes Strukturmodell für die aktive Stelle und den durch die fünf β–Stränge gebildeten Kernbereich des Proteins erzeugt. Durch die in der Vorlage fehlenden Bereiche (siehe Alignment in Abbildung 8.30) klafften in diesem Modell naturgemäß große Lücken, da es mit den heutigen Methoden nicht möglich ist, große Insertionen *ab initio* zu modellieren.

Es muß hier betont werden, daß – wie bei Blindvorhersagen im allgemeinen – niemals mit absoluter Sicherheit auf eine bestimmte Modellvorlage geschlossen werden kann. In diesem Fall haben sich mit der experimentellen Aufklärung der Struktur [45, 363] die meisten unserer Annahmen als richtig herausgestellt. Insbesondere die für die aktive Stelle des Proteins konnte ein Modell erzeugt werden,

für das 61 Aminosäurereste mit einer *RMS*–Abweichung von nur 1.41Å mit der experimentellen Struktur übereinstimmen. Abbildung 8.31 zeigt die zugehörige Superposition.

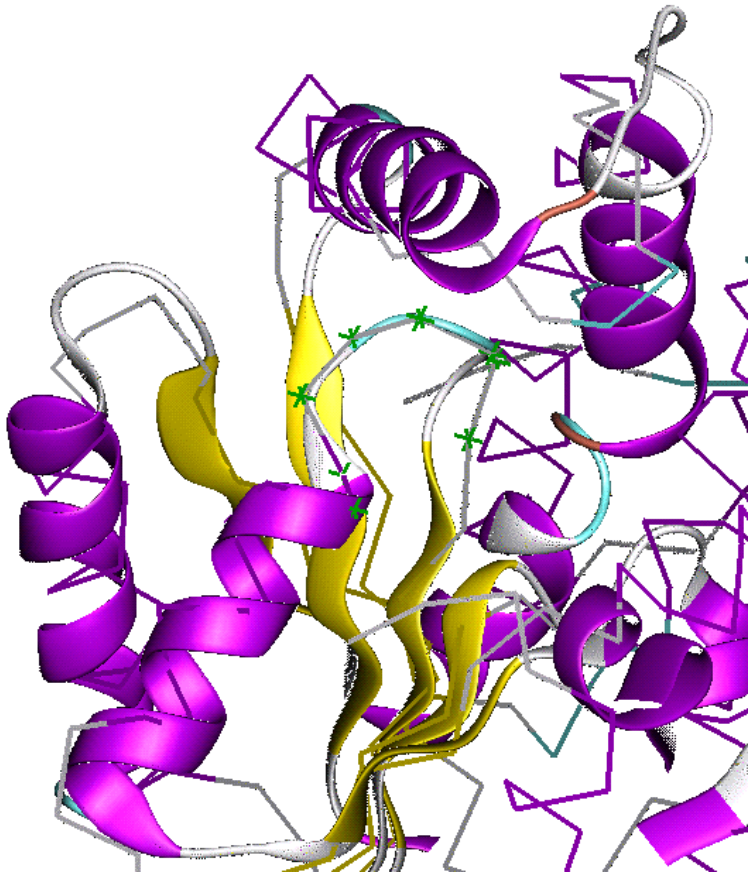


Abbildung 8.31: Aktive Stelle der TK: Superposition (61 Aminosäurereste, 1.41Å) von Modell (in der *Backbone*–Darstellung) und Struktur (in der *Cartoon*–Darstellung).

Eine Modellbildung für die großen Insertionen war – dem Paradigma der vergleichenden Modellierung entsprechend – nicht möglich, da entsprechende Vorlagen in der Strukturdatenbank nicht vorhanden waren. Zudem stellte sich durch die Strukturaufklärung heraus, daß die für die fehlenden Bereiche vorgesagten Helices eine große Dimerisierungsstelle ausbilden. Die Tatsache, daß die `TK` als Dimer vorliegt, war in der Kürze der zur Verfügung stehenden Zeit nicht vorherzusagen und auch aus schon länger betriebenen theoretischen Arbeiten anderer Gruppen zur Struktur nicht bekannt gewesen.
Ein struktureller Vergleich der experimentellen Struktur gegen die Datenbank der aufgeklärten Proteinstrukturen hat gezeigt, daß die von uns mit den Methoden der Faltungserkennung vorhergesagte Ähnlichkeit zu den Adenylatkinasen den Tatsachen entspricht und die Adenylatkinasen unter strukturellen wie funktionellen Kriterien die nächsten Verwandten der `TK` in der Datenbank sind [363].

Diese Ergebnisse zeigen, daß es trotz geringer Sequenz– und auch Strukturähnlichkeit in vielen Fällen mit Strukturvorhersagemethoden wie der RDP–Methode möglich ist, zumindest ein plausibles Modell für die aktive Stelle eines Proteins vorherzusagen, dessen dreidimensionale Struktur unbekannt ist.

8.7 Zusammenfassung der Ergebnisse

Dieses Kapitel dokumentiert den durch die RDP–Methode für die Proteinstrukturvorhersage erreichte Verbesserung durch

- Vergleich der Alignmentqualität gegen andere Methoden,
- Faltungserkennungsexperimente (ebenfalls im Vergleich) und
- echte Blindvorhersagen.

Abschnitt 8.3 vergleicht die Alignmentqualität der RDP–Methode gegen etablierte Methoden wie `Threader` und `123D` auf unterschiedlich schwierigen Testmengen. `123D` [5] ist eine Profilmethode (siehe Abschnitt 4.2.2), die keine Wechselwirkungspotentiale bei der Alignmentoptimierung einbezieht. `Threader` [168] benutzt dagegen wie die RDP–Methode Wechselwirkungspotentiale zur Berechnung von Sequenzstrukturalignments, verwendet jedoch ein anderes Optimierungsverfahren (siehe Abschnitt 4.2.3.2). Als Vergleichskriterium wird die Anzahl der Alignments, deren *RMS*–Abweichung geringfügig über der des als Referenz betrachteten Strukturalignments liegt, verwendet, da für die ausgewählten Testbeispiele nicht nur die Struktur, gegen die aliniert wird, sondern auch die Struktur der Sequenz bekannt ist.
Die Verbesserungen durch die RDP–Methode zeigen sich bereits auf der als am einfachsten eingestuften Testmenge (siehe Abschnitt 8.3.3). Für 201 von 205 Alignments bleibt das RDP–Alignment innerhalb der vorgegeben Toleranzgrenzen, während dies für `123D` nur in 184 und für `Threader` bestenfalls in 151 Fällen der Fall ist.
Dieser Trend bestätigt sich auch für die Testbeispiele, wo die Ähnlichkeit der Proteine bisher nur noch durch Strukturvergleichsmethoden aufgedeckt werden kann (siehe Abschnitt 8.3.5). Auch hier erreicht die RDP–Methode mit 45 von 73 Alignments eine wesentlich höhere Alignmentqualität als `123D` und `Threader`, die mit 24 beziehungsweise 20 Beispielen nur halb so viele gute Alignments liefern.
Trotz der hohen Alignmentqualität bleibt die Laufzeit der RDP–Methode im erträglichen Zeitrahmen (siehe Abschnitt 8.3.6). Die typische Laufzeit liegt unter einer Minute und auch für die schwierigen Fälle wird die Grenze von drei Minuten nur für ein Paar mit mehr als 500 Resten mit unter fünf Minuten geringfügig überschritten. Zum Vergleich dazu verweisen Lathrop und Smith auf Fälle, wo sie bereits für relativ kurze eindomänige Proteine ihre Methode trotz des von ihnen verwendeten einfacheren fragmentbasierten Sequenzstrukturalignmentmodells nach zwei Stunden erfolglos abbrechen mußten [195].

Die bessere Qualität der von der RDP–Methode berechneten Sequenzstrukturalignments wirkt sich auch positiv auf die Erkennung sehr entfernter struktureller Verwandtschaften aus. So erkennt die RDP–Methode in den in Abschnitt 8.4 vorgestellten, sehr schwierigen Erkennungsexperimenten in 57 % der Fälle ein Protein der gleichen Faltungsklasse auf Platz 1 der Rangliste (siehe Tabelle 8.9). Mit verschiedenen Austauschmatrizen erreicht man mit Sequenzalignment nur eine Erkennung von 33% (siehe Tabelle 8.6). Und mit `123D` findet man in 38% der Beispiele eine verwandte Faltung auf Rang 1 (siehe Tabelle 8.7).
Abschnitt 8.5 dokumentiert, daß es bereits mit dem zum damaligen Zeitpunkt vorhandenen Prototyp der RDP–Methode in mehreren Blindvorhersagen möglich war, den korrekten Faltungstyp zu identifizieren und auch relativ gute Sequenzstrukturalignments zu berechnen. So konnte im Konzert mit der `123D`–Methode in mindestens vier Fällen die korrekte Faltung identifiziert werden und in mindestens zwei Fällen ein als richtig anerkanntes Sequenzstrukturalignment eingereicht werden [217].
Im Vergleich dazu konnten mit der *Branch&Bound*–Methode von Lathrop und Smith [195] nur zwei Faltungen identifiziert und kein von den Evaluatoren als korrekt anerkanntes Alignment berechnet werden [217]. Dieses Beispiel zeigt, daß es zum Erfolg bei der Strukturvorhersage nicht unbedingt sinnvoll, die bezüglich einer empirischen Bewertungsfunktion und eines abstrakten Sequenzstrukturalignmentmodells optimale Lösung [217] zu berechnen, sondern vielmehr darauf ankommt, bei der Berechnung von Sequenzstrukturalignments signifikante lokale Ähnlichkeiten zu detektieren und zusätzliche biologischen und biochemischen Randbedingungen zu berücksichtigen, wie dies durch die RDP–Methode geschieht.
In der internen Nachanalyse des Wettbewerbs hat sich zudem herausgestellt, daß bei der Auswahl der eingereichten Alignments zu sehr auf die Kompaktheit des Alignments geachtet wurde. In zu kompakten Alignments werden auch die Teile von Sequenz und Struktur zusammengezwungen, die von den Kriterien der RDP–Methode korrekter Weise als unähnlich erkannt und damit eigentlich nicht aliniert werden.
Die mit der unter Einbeziehung der im `CASP II`–Wettbewerb gewonnen Erkenntnisse verbesserten RDP–Methode in der Nachanalyse der Wettbewerbsproteine erzielbaren Vorhersageergebnisse, die ebenfalls in Abschnitt 8.5 diskutiert werden, lassen für zukünftige Blindvorhersagen noch bessere Vorhersagen erwarten.
Bei der Blindvorhersagen für die Thymidinkinase des *Herpes Simplex Virus I* (siehe Abschnitt 8.6) hat sich gezeigt, daß die RDP–Methode durch die Einbeziehung biologischer Randbedingungen und die Strategie, lokale Ähnlichkeiten mit unterschiedlichen Bewertungskriterien zu suchen, Strukturmodelle erzeugt, die sich durch besondere Genauigkeit des Modells im Bereich der aktiven Stelle hervorheben, auch wenn sich die Ähnlichkeit zwischen der untersuchten Sequenz und der in der Datenbank vorhandenen ähnlichsten Struktur nur über Teilbereich der Sequenz erstreckt.

Kapitel 9

Ausblick und RDP–Erweiterungen

Die RDP–Methode soll nach Abschluß dieser Arbeit weiterentwickelt und auf genomische und Differenzprofil–Daten angewendet werden. Die Anwendung auf große Datenbestände erfordert, daß die Effizienz der RDP–Methode verbessert wird. Dies kann unter anderem dadurch erreicht werden, daß in Erkennungsexperimenten insbesondere bei den Paaren, die sich später als nicht ähnlich herausstellen, weniger Aufwand für die Berechnung des Sequenzstrukturalignments investiert wird. Zum einen ist es dazu notwendig, dies frühzeitig zu erkennen und durch geeignete Schranken oder Randbedingungen den Suchraum einzuschränken. Zum anderen muß dabei beachtet werden, daß eine Abschätzung der Signifikanz der Vorhersage über den *zscore* weiterhin möglich bleibt, obwohl die berechnete Bewertung einiger Alignments nicht mehr dem Optimum entspricht.
Die von der RDP–Methode verwendete Bewertungsfunktion ist empirisch und kann sicher weiter verbessert werden. Dazu muß zum einen die in Abschnitt 7.7 begonnene Kalibrierung der Gewichte der einzelnen Kostenfunktionsbestandteile auf eine breitere Datenbasis gestellt und auf Erkennungsexperimente ausgedehnt werden. Aber auch die einzelnen Kostenfunktionsbestandteile – sowohl die Wechselwirkungspotentiale als auch die `CCP`– und Hydrophobizitätspotentiale – müssen weiter verbessert werden. Zum Beispiel ist der Einfluß von familienspezifischen Potentialen und von auf der Basis multipler Alignments abgeleiteten Potentialen auf die Alignmentqualität zu untersuchen.
Die Alignmentqualität soll aber nicht nur durch die Verbesserung der Bewertungsfunktion, sondern auch durch die folgenden Erweiterungen und Verbesserungen der RDP–Methode weiter erhöht werden:

- Die Modellierung von Schleifen ist eines der nicht zufriedenstellend gelösten Probleme in der vergleichenden Modellierung, insbesondere dann, wenn es sich nicht um kurze Schleifen mit wenigen Resten handelt. Die Ursachen dafür liegen häufig in der *ab initio*–Modellierung der Schleifen und in Fehlern, die bereits aus dem berechneten Alignment resultieren. Abschnitt 9.1 skizziert einen Ansatz, wie dieses Problem durch Integration der Schleifenmodellierung in die Alignmentberechnung mittels RDP gelöst werden kann.

- Die Einbeziehung verwandter Sequenzen ist bei der Berechnung von Alignments eine anerkannte Methode die Alignment– und Vorhersagequalität zu steigern (siehe zum Beispiel Abschnitt 4.3.1). Abschnitt 9.2 skizziert, wie die RDP–Methode zu einer multiplen Sequenzstrukturalignmentmethode erweitert werden kann.

- Biologisches Wissen und die Kenntnisse des Anwenders einer Vorhersagemethode über die von ihm untersuchte Sequenz werden bisher in Sequenzstrukturalignmentmethoden zu wenig genutzt. In Abschnitt 9.3 wird die Erweiterung der RDP–Methode um ein Regelsystem vorgeschlagen.

Neben der weiteren Verbesserung der RDP–Sequenzstrukturalignmentmethodik soll im Anschluß an diese Arbeit die Anwendung der RDP–Methode auf andere bioinformatische Vergleichsprobleme untersucht werden (siehe Abschnitte 9.4 bis 9.6).

9.1 Verknüpfung von Sequenzstrukturalignment und Schleifenmodellierung

In der heute üblichen Vorgehensweise der vergleichenden Modellierung werden die einzelnen in Abschnitt 4.1 beschriebenen Schritte sequentiell ausgeführt. Dies hat häufig zur Folge, daß bereits beim Schließen des Rückgrates des aus dem Alignment resultierenden Modells während der Schleifenmodellierung Probleme auftreten, die entweder große Umlagerungen von alinierten Resten oder nicht zusammenhängende Modelle zur Folge haben. Die Ursache dafür liegt darin, daß die Abbildung der Sequenz auf die Modellstruktur nicht in allen Teilen korrekt ist und der Aspekt der Modellierbarkeit der Schleifen (unalinierte Bereiche in Sequenz und Struktur) in der Regel nicht von vorneherein berücksichtigt wird.
Die RDP–Methode bezieht dagegen die Modellierbarkeit von Schleifen auch bei Faltungerkennungsexperimenten bereits in die Alignmentberechnung mit ein. So werden Kriterien für die Modellierbarkeit von Schleifen sowohl in der *top–down*–Phase als auch in der *bottom–up*–Phase der RDP–Methode zur Auswahl und Modifikation zulässiger Sequenzstrukturalignments verwendet (siehe Abschnitte 7.3 und 7.6.2). In der bisherigen Implementierung werden jedoch nur einfache Kriterien eingesetzt:

- Bei Deletionen bezüglich der Struktur wird über einfache Abstandsbedingungen sichergestellt, daß die Schleife durch die Reste der Sequenz theoretisch schließbar ist, ohne die Lage der alinierten Reste zu verändern.
- Bei Insertionen bezüglich der Struktur wird getestet, ob eine Insertion an der durch das Alignment vorgeschlagenen Stelle möglich ist, ohne die Lage der alinierten Reste stark zu verändern. Dazu wird anhand der Zugänglichkeit der an die Insertion angrenzenden alinierten Positionen überprüft, ob die Proteinkette an diesen Stellen in das Lösungsmittel ausweichen kann.

In beiden Fällen wird das Alignment so modifiziert, daß die Kriterien erfüllt sind, oder verworfen, wenn dies nicht möglich ist. So ist sichergestellt, daß das von der RDP–Methode berechnete Strukturmodell in ein Modell mit geschlossener Peptidkette überführt werden kann. Die RDP–Methode ist darauf ausgelegt, weitere

Kriterien für die Zulässigkeit von Alignments einzubeziehen, wie zum Beispiel das Verbot von Insertionen und Deletionen in den aktiven Zentren und zusätzliche vom Anwender spezifizierte Abstandsbedingungen.

Dies stellt jedoch nur einen ersten Versuch dar, das Schleifenmodellierungsproblem bereits bei der Berechnung der allen weiteren Schritten zugrundeliegenden Abbildung der Sequenz in die Struktur zu berücksichtigen, und nutzt die Möglichkeiten der RDP–Methode bei weitem nicht aus.

In jedem Stadium der Berechnung des Alignments mit der RDP–Methode liegt Information über Teile des letztendlichen Strukturmodells vor, da bereits Teile des Alignments berechnet wurden. Diese Information soll in Zukunft genutzt werden, um bereits bei der Berechnung von Teilalignments strukturelle Vorschläge für die Schleifen zu berechnen, die nicht aus der zugrundeliegenden Struktur übernommen werden können, sondern neu zu modellieren sind.

Die Idee dieser Schleifenmodellierung mit RDP ist, die Schleifen so zu modellieren, daß die Aminosäurereste der Schleifen optimal mit dem bisher berechneten Strukturmodell wechselwirken, wobei auch die Lösungsmittelzugänglichkeit der dadurch eventuell vergrabenen Reste zu berücksichtigen ist. Dazu soll – ähnlich wie bei der Berechnung neuer Teillösungen unter Einbeziehung des Wechselwirkungspotentials (siehe Abschnitt 7.2.3 und Abbildung 7.5) – das von den bereits alinierten Positionen induzierte Wechselwirkungspotential ausgenutzt werden, um ein günstiges Modell für die zu modellierende Schleife zu berechnen.

Da die hier betrachteten Wechselwirkungspotentiale diskrete Wechselwirkungen zwischen Aminosäuren bewerten und diese Potentiale nur eine eingeschränkte Auflösung haben, kann die Suche nach einem Schleifenmodell als Wegeproblem in einem diskreten Gitter zwischen den durch die Schleife zu verbindenden Positionen im bisherigen Strukturmodell formuliert werden. Dabei wird die Belegung eines Gitterknotens mit einem Aminosäurerest durch die Wechselwirkung des Aminosäuretyps mit den bereits alinierten Strukturpositionen bewertet.

Sterische Randbedingungen können dabei durch bereits belegte Gitterknoten einfach modelliert werden. Die Länge eines möglichen Weges ist durch die Anzahl der für die Schleife zur Verfügung stehenden Aminosäurereste vorgegeben. Existiert kein Weg dieser Länge, so kann die dem Strukturmodell zugrundeliegende Teillösung nicht zu einem korrekten Strukturmodell erweitert werden und muß entweder modifiziert oder verworfen werden. Existiert ein Weg, so kann das zugehörige Strukturmodell nicht nur in der anschließenden vergleichenden Modellierung weiter verwendet werden, sondern die Wechselwirkungsenergie des Schleifenmodells kann auch von der RDP–Methode bei der Bewertung und Auswahl von Teillösungen verwendet werden.

Die Integration von Sequenzstrukturalignment und Schleifenmodellierung in der RDP–Methode hat das Potential, sowohl die Qualität der Proteinstrukturvorhersage als ganzes, als auch die Qualität und Berechnung des Sequenzstrukturalignments im einzelnen weiter zu verbessern:

- Das von der RDP–Methode berechnete Sequenzstrukturalignment kann in jeden Falle in ein vollständiges Strukturmodell umgesetzt werden.
- Neben der Abbildung der Sequenz in die unterliegende Modellstruktur liefert die RDP–Methode auch erste Lösungsvorschläge für die nicht abgebildeten Bereiche.
- Teillösungen, die nicht zu einem vollständigen Strukturmodell führen können, können verworfen werden und müssen nicht weiterverfolgt werden.
- Bei der Bewertung von Teillösungen können nicht nur die alinierten Aminosäurereste, sondern auch die Reste in neu zu modellierenden Schleifen berücksichtigt werden.

9.2 Multiples RDP–Sequenzstrukturalignment

Die RDP–Methode eignet sich auch zur Berechnung multipler Sequenzstrukturalignments. Multiple Sequenzen können methodisch bereits in der Form eines voralinierten Alignmentprofils durch die RDP–Methode auf eine Struktur abgebildet werden. Es ist jedoch noch näher zu untersuchen, wie die Abbildung eines Sequenzvektors auf eine Strukturposition zu *bewerten* ist und welchen Einfluß die Variabilität der Sequenzen des Profils untereinander auf das Ergebnis haben. Im nächsten Schritt soll die RDP–Methode so ausgebaut werden, daß auch das Alignment der zusätzlichen Sequenzen zusammen mit der Abbildung auf die Struktur berechnet wird. Dazu sollen Baumheuristiken eingesetzt werden und der für das multiple Alignment vorgeschlagene Lösungsansatz (siehe Abschnitt 9.5) auf das Sequenzstrukturalignment mit mehreren Sequenzen übertragen werden.
Schwieriger wird es, wenn nicht nur mehrere Sequenzen, sondern auch mehrere Strukturen bei der Berechnung des multiplen Alignments verwendet werden sollen. Sind mehrere Strukturen zu einem Faltungstyp bekannt, könnte zum Beispiel das Alignment gegen den gemeinsam Strukturkern oder auch gegen eine vorabberechnete Konsensusstruktur berechnet werden. Es ist aber auch denkbar, daß durch das Alignment von Sequenzen gegen mehrere Strukturen ein Strukturmodell entsteht, das aus Fragmenten aus den unterschiedlichen Strukturen zusammengesetzt ist.

9.3 Regelbasierte Steuerung der RDP–Methode

Die in dieser Dissertation entwickelte RDP–Methode erlaubt es, biologische, beziehungsweise biochemische Information, sowohl in der Form von Randbedingungen als auch als Bestandteil der Kostenfunktion in den Berechnungsprozeß einzubeziehen. Innerhalb eines beantragten DFG–Projektes soll ein System entwickelt werden, das es erlaubt, biologisches Wissen explizit in einer entsprechenden Beschreibungssprache zu formulieren und systemgestützt in effiziente Optimierungs-

verfahren wie die RDP–Methode umzusetzen. Durch dieses System sollen Anwender biologische Fakten in Alignmentbewertungssyteme einbringen und auf hohem Niveau steuern können.
Desweiteren soll die Beschreibungssprache zur automatischen Formulierung von Regeln für das Sequenzstrukturalignment aus Datenbanken und experimentellen Messungen eingesetzt werden. Derzeit werden biologisches Wissen und experimentelle Daten nur implizit, zum Beispiel in der eigentlichen Berechnung nachgeschalteten Filterprozeduren genutzt. Der monolithischen Aufbau der meisten Sequenzstrukturalignmentalgorithmen erlaubt es in der Regel nicht, zusätzliche Randbedingungen bei der Berechnung effektiv und einfach zu berücksichtigen. Die RDP–Methode dagegen kann aufgrund ihres modularen Aufbaus sehr einfach um Module erweitert werden, die zum Beispiel Teillösungen auf die Erfüllung von Regeln aus experimentellen Daten hin überprüfen. Erste Beispiele dafür wurden bereits in Kapitel 7 diskutiert.

9.4 Sequenzalignment mit RDP

Die RDP–Methode kann auch zur Berechnung von reinen Sequenzalignments verwendet werden. Dazu sind nur die Kostenfunktionsterme und Filterfunktionen für Teillösungen auszuschalten, die Strukturinformation in die Berechnung einbeziehen. Der Vorteil der RDP–Methode beim Sequenzalignment besteht darin, daß Insertionen und Deletionen automatisch aufgrund der Art der Berechnung als Bereiche übrig bleiben, die nicht mehr aufgrund signifikanter Ähnlichkeiten zugeordnet werden können. Daher müssen keine expliziten Kosten für Insertionen und Deletionen spezifiziert werden und das Problem der Gapkosten wird somit in natürlicher Weise umgangen. Als Orakel können dazu sowohl das lokale Sequenzalignment als auch die von Programmen wie `BLAST` oder in Datenbanken wie `PROSITE` gefundenen lokalen Ähnlichkeiten verwendet werden.

9.5 Multiples Sequenzalignment mit RDP

Multiples Alignment kann schon auf Sequenzebene signifikant zur Alignmentqualität gemessen in Abweichung von Strukturalignment beitragen [125]. Angeregt durch die RDP–Methode [340] wurden im `PROTAL`–Projekt bereits zwei multiple Alignmentmethoden entwickelt, die ebenfalls auf dem *divide&conquer*–Prinzip beruhen [235, 329, 342]. In [329, 342] wird rekursiv jeweils eine der Sequenzen in der Mitte aufgetrennt und dann mittels paarweisem Sequenzalignment nach günstigen Aufteilungspositionen in den anderen Sequenzen gesucht.
In [235] wird eine Vorgehensweise verfolgt, die sich noch näher am RDP–Ansatz orientiert. Hier werden der generellen Idee der RDP–Methode folgend zunächst paarweise die Sequenzabschnitte einander zugeordnet, die eindeutig oder mit hoher Zuverlässigkeit zuzuordnen sind, und die so bestimmten Paare werden im

nächsten Schritt mit einer *Greedy*–Methode zu einem multiplen Alignment zusammengesetzt. Die *Greedy*–Vorgehensweise ist beim Zusammensetzen multipler Alignments aber keine gute Wahl, da dadurch die Güte der Lösung stark vom Kriterium abhängt, das zur Festlegung der Reihenfolge zur Integration der Teillösungen in Gesamtlösungen dient. Durch die RDP–Methode wird dieses Problem auf natürliche Weise umgangen, da die RDP–Methode viele alternative Teillösungen und damit auch unterschiedliche Gesamtlösungen gleichzeitig verfolgt, und somit Fehlern entgegenwirkt, die durch die Reihenfolgeabhängigkeit entstehen.
Die geplante Erweiterung der RDP–Methode auf multiples Sequenzstrukturalignment (siehe Abschnitt 9.2) ermöglicht aufgrund der Modularität der Methode die direkte Anwendung von RDP auf das multiple Sequenzalignmentproblem.

9.6 Strukturalignment mit RDP

Wie bereits in Abschnitt 6.1 erwähnt, kann die RDP–Methode auch für andere schwierige bioinformatische Vergleichsprobleme, wie zum Beispiel das Strukturalignment, verwendet werden. Dazu sind im wesentlichen die verwendete Kostenfunktion und die Funktionen anzupassen, die Teillösungen generieren beziehungsweise aussortieren.
In einem ersten Prototyp sind bereits Verfahren implementiert, lokale strukturelle Ähnlichkeiten zu finden. Eine Möglichkeit besteht darin, alle Fragmente einer festvorgegebenen Länge aus den zu vergleichenden Strukturen optimal zu superpositionieren und dann der RDP–Methode die Suche nach einer maximalen Menge von gut kombinierbaren Fragmenten zu überlassen.
Es ist jedoch auch möglich, nach lokalen Strukturähnlichkeiten mittels lokalem Alignment der ϕ/ψ–Winkelbeschreibungen der zu vergleichenden Proteinstrukturen mit geeigneten Parametern zu suchen. Allgemein können die ϕ/ψ–Winkel nicht als Maß für das Strukturalignment dienen, dessen Ziel ein Alignment mit einer möglichst geringen *RMS*–Abweichung ist. Lokale Ähnlichkeiten können jedoch durch dynamische Programmierung aufgedeckt werden, da zwei Fragmente auch gut superpositionieren, wenn alle ϕ/ψ–Winkel sehr ähnlich sind und keine Insertionen oder Deletionen in dem zugehörigen Alignment enthalten sind (die Umkehrfolgerung gilt in der Regel nicht). Die Verträglichkeit von so gefundenen Fragmentzuordnungen wird durch entsprechende Filterfunktionen beim rekursiven Zusammensetzen der Gesamtlösung sichergestellt.
Durch Relaxation der Bedingung der üblicherweise benutzten *Rigid body*–Superpositionierbarkeit ist es möglich, mit der RDP–Methode neue, bisher nicht erkannte strukturelle Ähnlichkeiten zwischen Proteinstrukturen aufzudecken. Diese Ähnlichkeiten sind insbesondere zur Erklärung und Interpretation von Ergebnissen aus Faltungserkennungsexperimenten sehr aufschlußreich, da dabei nicht immer eine Ähnlichkeit erkannt wird, die aus einer guten Superpositionierbarkeit der Gesamtstrukturen, sondern aus strukturellen Ähnlichkeiten von Faltungsbereichen resultiert.

Kapitel 10

Zusammenfassung

Diese Arbeit befaßt sich mit Algorithmen für eines der zentralen Probleme der molekularen Bioinformatik [199, 201], der theoretischen Proteinstrukturvorhersage. Der entwickelte Lösungsansatz stützt sich dabei methodisch auf die ähnlichkeitsbasierte Proteinstrukturvorhersage ab, die gegenwärtig den einzigen in vielen Fällen erfolgreichen Weg zur Vorhersage der dreidimensionalen Struktur von Proteinen darstellt.
Der Schwerpunkt dieser Dissertation liegt auf der Entwicklung von Algorithmen und Bewertungssystemen zur Abbildung von Sequenzen auf experimentell aufgeklärte Proteinstrukturen, allgemein als Sequenzstrukturalignment oder auch *Threading* bezeichnet. Die Berechnung dieser Abbildung stellt den ersten und damit grundlegenden Arbeitsschritt der ähnlichkeitsbasierten Proteinstrukturvorhersage dar. Fehler, die hier gemacht werden, können in den nachfolgenden Schritten (siehe Abschnitt 4.1) nur schwer und sehr häufig gar nicht mehr behoben werden.
Die in dieser Arbeit entwickelte Methode der Rekursiven Dynamischen Programmierung (RDP) zielt in erster Linie auf die Berechnung strukturrichtiger Sequenzstrukturalignments, die sowohl zur Vorhersage von strukturellen Verwandtschaften als auch als Startpunkt für die vergleichende Modellierung dienen. Das algorithmische Grundgerüst der RDP–Methode ist ein modular aufgebauter *divide & conquer*–Algorithmus, der auch für andere bioinformatische Vergleichsprobleme, wie das Proteinstrukturalignment und das multiple Sequenzalignment, angewendet werden kann (siehe Kapitel 9).
Im Unterschied zu anderen, vom Grundkonzept ähnlichen Verfahren [195, 369], die das fragmentbasierte Sequenzstrukturalignmentproblem adressieren (siehe Abschnitt 4.2.1), verwendet die RDP–Methode [340] den *divide & conquer*–Ansatz zur Lösung des allgemeinen, nicht fragmentbasierten Sequenzstrukturalignmentproblems.
Bei der Berechnung strukturrichtiger Alignments kommen unterschiedliche Bewertungssysteme zum Einsatz (siehe Kapitel 5), angefangen von einfachen Aminosäureaustauschmatrizen bis hin zu komplexen Pseudoenergiepotentialen, die Wechselwirkungen zwischen zwei oder mehr Aminosäureresten in einer dreidimensionalen Struktur bewerten und aufgrund ihrer Nichtlokalität das Sequenzstrukturalignmentproblem zu einem *NP*–vollständigen Problem machen [193]. Gemeinsam ist allen diesen Bewertungssystemen jedoch, daß sie mit statistischen Methoden aus diskretisierten Beschreibungen der biologischen und biochemischen Realität abgeleitet sind.
Sind die zu alinierenden Proteine evolutionär eng verwandt, so sind diese Bewer-

tungssysteme hinreichend genau und das bezüglich des Bewertungssystems optimale Alignment stimmt weitgehend mit dem strukturrichtigen Alignment überein. Diese Situation ändert sich, sobald mit den Methoden des Sequenzstrukturalignments entferntere, rein strukturelle Ähnlichkeiten aufgedeckt werden sollen, die auf der Ebene der Proteinsequenz nicht mehr nachweisbar sind. In diesem Falle ist ein Ähnlichkeitsmaß zwischen einer Sequenz und einer Struktur gefragt, welches die gesuchten strukturellen Ähnlichkeiten erkennt, ohne jedoch zu detailgetreu zu sei, da zwar die allgemeine Faltung gleich, die Struktur jedoch aufgrund der unterschiedlichen Aminosäurereste im Detail unterschiedlich ist.
Daher kommt es darauf an, bei der Beschreibung einer Proteinfaltung durch Wechselwirkungen zwischen Strukturpositionen und der Bewertung dieser Wechselwirkungen durch aminosäuretypabhängige Wechselwirkungspotentiale ein geeignetes Abstraktionsniveau zu wählen. Die in der RDP–Methode verwendeten, auf Voronoikontaktrelationen basierenden Potentiale realisieren dieses Abstraktionsniveau, indem Wechselwirkungen nur zwischen räumlich direkt benachbarten Aminosäuren berücksichtigt werden und Wechselwirkungen im Unterschied zu anderen empirischen Potentialen (siehe Abschnitt 5.2.4) nur sehr grob bezüglich ihres Wechselwirkungsabstands klassifiziert werden.
Wechselwirkungspotentiale reichen alleine allerdings nicht aus, um zwischen guten und schlechten Sequenzstrukturalignments zu unterscheiden (siehe Abschnitt 5.2.5 und Diplomarbeit Halfmann [131]). Aus diesem Grund verwendet die RDP–Methode im Unterschied zu vielen anderen Sequenzstrukturalignmentverfahren [48, 168, 195, 321] eine gemischte Bewertungsfunktion, die neben Anteilen, die die Sequenzähnlichkeit bewerten, auch das Kontaktkapazitätspotential `CCP` [5] enthält, welches im wesentlichen den Hydrophobizitätsanteil enthält, der in Wechselwirkungspotentialen unterrepräsentiert ist. Die Strukturvorhersage für die Nukleotidyltransferase im Rahmen des `CASP II`–Wettbewerbs (siehe Abschnitt 8.5.1) ist ein gutes Beispiel gegen die verbreitete These, daß die Sequenzähnlichkeit beim Sequenzstrukturalignment von strukturell ähnlichen Proteinen nicht nutzbar ist.
Die RDP–Methode hebt sich außerdem von anderen Methoden dadurch ab, daß sie aufgrund der im Sequenzstrukturalignmentzusammenhang bekannten Schwächen globaler Bewertungsfunktionen nicht nach optimalen, sondern nach Lösungen mit guter Bewertung sucht, die zusätzliche Kriterien und Randbedingungen erfüllen. Zum Beispiel sind die aktiven Zentren zwischen verwandten Proteinen in vielen Fällen in ihrer Aminosäuresequenz wesentlich stärker konserviert als dies für die Gesamtsequenzen der Fall ist. Ein berechnetes Modell sollte dies in jedem Falle wiedergeben. Die Idee, rekursiv signifikante Teillösungen mit sich adaptiv anpassenden Bewertungsfunktionsbestandteilen zu suchen, nutzt dieses biologische Fakt aus und führt auch dann zu aussagekräftigen Modellen des aktiven Zentrums, wenn sich – wie im Beispiel der vorhergesagten Thymidinkinasestruktur (siehe Abschnitt 8.6) – die strukturelle Ähnlichkeit auf Teilstrukturbereiche beschränkt. Als Nebeneffekt kann der Anteil des durchlaufenen Suchraums durch die Einschränkung auf signifikante Teillösungen stark reduziert werden, obwohl

effizient berechenbare Schranken für die Bewertungsfunktion fehlen.
Die RDP–Methode berechnet für ein Teilproblem im rekursiven Abstieg nicht nur eine optimale Lösung sondern eine Menge von Lösungen, die optimale oder auch suboptimale Lösungen bezüglich unterschiedlicher Bewertungssysteme sind. Damit begegnet die RDP–Methode dem für *divide & conquer*–Heuristiken typischen Problem, daß in Stadien des Verfahrens, wo noch wenig Information über das schließlich berechnete Gesamtmodell vorliegt, Fehlentscheidungen getroffen werden, die im weiteren Verlauf nicht mehr behoben werden können.
Im Unterschied zu anderen Sequenzstrukturalignmentmethoden garantiert die RDP–Methode durch die explizite Einbeziehung von Randbedingungen für die Modellierbarkeit von Schleifen in den Berechnungsprozeß, daß ein berechnetes Sequenzstrukturalignment immer in ein zulässiges Strukturmodell mit geschlossenem Rückgrat übersetzt werden kann. Teillösungen, die Insertionen oder Deletionen enthalten, die nicht zu einem guten Strukturmodell führen können, werden von der RDP–Methode möglichst frühzeitig verworfen. Teillösungen, die nur unter Anpassung der an die Insertionen oder Deletionen angrenzenden Positionen zu einem sinnvollen Strukturmodell führen können, werden entsprechend modifiziert, um danach einer Neubewertung unterzogen zu werden (siehe Abschnitte 7.3 und 7.6.2).
Da die RDP–Methode dann stoppt, wenn keiner der verschiedenen Bewertungsfunktionsbestandteile signifikante Ähnlichkeiten zwischen noch nicht zugeordneten Sequenz– und Strukturbereichen zu Tage fördert, bleiben diese Bereiche automatisch als Insertionen und Deletionen zurück. Diese Vorgehensweise hat im wesentlichen zwei Vorteile:

- Bereiche, für deren Ähnlichkeit keine der in den Orakel (siehe Abschnitt 7.2) verwendeten Bewertungsfunktionen eine signifikante Unterstützung nachweist, werden nicht aliniert. Dies zeigt sich nicht nur an den niedrigeren *RMS*–Abweichungen für die in Abschnitt 8.3 berechneten Sequenzstrukturalignments mit Sequenzen mit bekannter Struktur, sondern ist auch für die nachfolgenden Strukturvorhersageschritte sehr wichtig, da bei der Modellierung durch ein Alignment fälschlicherweise suggerierte Ähnlichkeiten sehr störend sind.

- Die RDP–Methode benötigt keine besonderen Bestrafungsterme für Insertionen und Deletionen (Gapkosten), um strukturrichtige Sequenzstrukturalignments zu berechnen. Diese methodische Eigenschaft von RDP ist von besonderem Vorteil, da die geeignete Wahl von Gapkosten in allen bisher vorgestellten Methoden – sei es zum Sequenzalignment oder Sequenzstrukturalignment – problematisch ist. Da die Bestrafung von Insertionen und Deletionen nicht zur Berechnung strukturrichtiger Alignments notwendig sind, können derartige Bestrafungsterme als orthogonales Kriterium bei der Unterscheidung zwischen verwandten und nicht verwandten Proteinen in Erkennungsexperimenten eingesetzt werden.

Wie gut die verschiedenen in der RDP–Methode verwendeten Heuristiken ihre Aufgaben erfüllen, zeigen die in Kapitel 8 diskutierten Ergebnisse. Im Vergleich der RDP–Methode gegen etablierte Methoden wie `Threader` und `123D` konnte die Qualität der Sequenzstrukturalignments wesentlich verbessert werden. So konnte zum Beispiel mit der RDP–Methode auf der schwierigsten Testmenge von Paaren mit rein struktureller Ähnlichkeit die Anzahl guter Alignments gegenüber den anderen Methoden (bei mit `Threader` vergleichbaren Laufzeiten) fast verdoppelt werden (siehe Abschnitt 8.3).
Die bessere Qualität der von der RDP–Methode berechneten Sequenzstrukturalignments wirkt sich auch positiv auf die Erkennung sehr entfernter struktureller Verwandtschaften aus. So erkennt die RDP–Methode in den in Abschnitt 8.4 vorgestellten, sehr schwierigen Erkennungsexperimenten in 57 % der Beispiele ein Protein der gleichen Faltungsklasse auf Platz 1 der Rangliste (Sequenzalignment 33%, `123D` 38%).
Zudem hat die RDP–Methode in echten Blindvorhersagen für die Thymidinkinase des *Herpes Simplex Virus I* (siehe Abschnitt 8.6) und im Rahmen des `CASP II`–Strukturvorhersagewettbewerbs (siehe Abschnitt 8.5) ihre Qualitäten unter Beweis gestellt. Die ebenfalls in Abschnitt 8.5 diskutierten mit der nach `CASP II` verbesserten RDP–Methode erzielbaren Vorhersageergebnisse stimmen optimistisch für den gerade beginnenden `CASP III`–Wettbewerb.
Die RDP–Methode wird bereits seit geraumer Zeit über das `WWW` tagtäglich von Benutzern aus aller Welt zur Bearbeitung von Strukturvorhersageproblemem erfolgreich eingesetzt.
Durch die in dieser Arbeit entwickelten Methoden konnte sowohl der Anwendungsbereich der theoretischen Proteinstrukturvorhersage ausgedehnt als auch die Qualität der Vorhersagen wesentlich gesteigert werden. Die in Kapitel 9 skizzierten Erweiterungen und Verbesserungen verfolgen die Idee, daß insbesondere durch die Hinzunahme von mehr biologischen Randbedingungen in früheren Phasen des Algorithmus und zusätzlicher evolutionärer Information (zum Beispiel in Form verwandter Sequenzen) weitere Fortschritte erzielbar sind.
Es besteht die Hoffnung, daß die Anwendung der RDP–Methode auf die anstehenden großen Datenmengen aus der Genomsequenzierung und aus Expressionsexperimenten einen wichtigen Beitrag zu so wissenschaftlich interessanten und für die pharmazeutische Industrie auch kommerziell wichtigen Fragestellungen liefern kann, wie sie die Identifizierung von Zielproteinen (*targets*) für mittels rationalem Wirkstoffentwurf zu entwickelnde, neuartige Arzneistoffe darstellt.

Literaturverzeichnis

[1] R.A. Abagyan and A. Batalov. Do aligned sequences share the same fold? *Journal of Molecular Biology*, 273:355–368, 1997.

[2] T. Akutsu. On pattern matching methods for three dimensional protein structures. In *Proceedings of genome informatics workshop* `IV`, volume 88, pages 1–9, 1993.

[3] V. Alesker, R. Nussinov, and H.J. Wolfson. Detection of non–topological motifs in protein structures. *Protein Engineering*, 9(12):1103–1119, 1996.

[4] N. Alexandrov and N. Go. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Science*, 3:866–875, 1994.

[5] N. Alexandrov, R. Nussinov, and R. Zimmer. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In Lawrence Hunter and Teri E. Klein, editors, *Pacific Symposium on Biocomputing'96*, pages 53–72, PO Box 128, Farrer Road, Singapore 912805, 1996. World Scientific Publishing Co.

[6] N. N. Alexandrov. SARFing the PDB. *Protein Engineering*, 9(9):727–732, 1996.

[7] N. N. Alexandrov and D. Fischer. Analysis of topological and nontopological structural similarities in the PDB: New examples with old structures. *PROTEINS: Structure, Function and Genetics*, 25:354–365, 1996.

[8] N. N. Alexandrov, K. Takahashi, and N. Go. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *Journal of Molecular Biology*, 225:5–9, 1992.

[9] S.F. Altschul and W. Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 1996.

[10] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[11] D.E Anderson, W.J. Becktel, and F.W. Dahlquist. pH–induced denaturation of proteins: A single salt bridge contributes $3-5 kcal/mol$ to the free energy of folding T4 lysozyme. *Biochemistry*, 29:2403–2408, 1990.

[12] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[13] C.B. Anfinsen, E. Haber, M. Sela, and F.H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences USA*, 47:1309–1314, 1961.

[14] K. Arnheim *et al. Lexikon der Biochemie und Molekularbiologie*. Spektrum Akademischer Verlag, Heidelberg, 1995.

[15] A. Aszódi, M.J. Gradwell, and W.R. Taylor. Global fold determination from a small number of distance restraints. *Journal of Molecular Biology*, 251:308–326, 1995.

[16] A. Aszódi and W.R. Taylor. Homology modelling by distance geometry. *Folding & Design*, 1:325–334, 1996.

[17] A. Bairoch. PROSITE a dictionary of protein sites and patterns. Technical report, EMBL BIOcomputing, 1989.

[18] A. Bairoch. The ENZYME data bank. *Nucleic Acid Research*, 21:3155–3156, 1993.

[19] A. Bairoch and B. Boeckmann. The SwissProt protein sequence data bank. *Nucleic Acid Research*, 20:2019–2022, 1992.

[20] A. Bairoch, P. Bucher, and K. Hofmann. The PROSITE database, its status in 1995. *Nucleic Acid Research*, 24(1):189–196, 1996.

[21] R.L. Baldwin. Matching speed and stability. *Nature*, 369:183–184, 1994.

[22] C.B. Barber, D.P. Dobkin, and H. Huhdanpaa. The `quickhull` algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.

[23] G.J. Barton. Protein secondary structure prediction. *Current Opinion in Structural Biology*, 5(5):372–376, 1995.

[24] A. Bauer and A. Beyer. An improved pair potential to recognize native protein folds. *PROTEINS: Structure, Function and Genetics*, 18:254–261, 1994.

[25] H. Bekker, H.J.C. Berendsen, and W.F. van Gunsteren. Force and virial of torsional-angle-dependent potentials. *Journal of Computational Chemistry*, 16(5):527–533, 1995.

[26] S.A. Benner, I. Badcoe, M.A. Cohen, and D.L. Gerloff. Bona fide prediction of aspects of protein conformations: Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. *Journal of Molecular Biology*, 235:926–958, 1994.

[27] S.A. Benner and D. Gerloff. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Adv. Enzym. Regul.*, 31:121–181, 1991.

[28] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP–complete. *Journal of Computational Biology*, 5(1):27–40, 1998.

[29] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Jr. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: a computer based archival file for macromolecular structures. *Journal of Molecular Biology*, 112:535–542, 1977.

[30] D.L. Beveridge and F.M. DiCapua. Free energy via molecular simulation: A primer. In W.F. van Gunsteren and P.K. Weiner, editors, *Computer simulation of biomolecular systems*, pages 1–26. Alliant Computer Systems Corporation, ESCOM, 1989.

[31] H. Bielka, H. Dixon, P. Karlson, N. Liebécq, C.and Sharon, E. Van Lenten, S. Velick, J. Vliegenthart, and E. Webb. *Enzyme Nomenclature.* Academic Press Inc., London, 1984.

[32] F.R. Blattner *et al.* The complete genome sequence of *escherichia coli* K-12. *Science*, 277:1453–1462, 1997.

[33] T.L. Blundell, B.L. Sibanda, M.J. Sternberg, and J.M. Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326(6111):347–352, 1987.

[34] E.M. Boczko and C.L. Brooks III. First–principles calculation of the folding free energy of a three–helix bundle protein. *Sciene*, 269:393–215, 1995.

[35] H.-J. Böhm. Site-directed structure generation by fragment-joining. *Perspectives in Drug Discovery and Design*, 3:21–33, 1995.

[36] H.-J. Böhm, G. Klebe, and H. Kubinyi. *Wirkstoffdesign.* Spektrum Akademischer Verlag, 1996.

[37] P. Bork and E.V. Koonin. Predicting functions from protein sequences – where are the bottlenecks? *Nature Genetics*, 18:313–318, 1998.

[38] S. Bouaziz, C. van Heijenoort, J.-C. Huet, J.-C. Pernollet, and E. Guittet. ^{1} H and ^{15} N resonanze assignment and secondary structure of capsicein, an α–elicitin by three-dimensional heteronuclear NMR. *Biochemistry*, 33:8188–8197, 1994.

[39] J.U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.

[40] C. Branden and J. Tooze. *Introduction to Protein Structure.* Garland Publishing Inc., New York, N. Y., 1991.

[41] S.E. Brenner and A. Berry. Protein design by optimization of a sequence–structure quality funktion. In *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, pages 44–52. AAAI/MIT, 1994.

[42] S.E. Brenner, C. Chothia, and T.J.P. Hubbard. Population statistics of protein structures: lessons from structural classfication. *Current Opinion in Structural Biology*, 7:369–376, 1997.

[43] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[44] C.L. Brooks III. Methodological advances in molecular dynamics simulations of biological systems. *Current Opinion in Structural Biology*, 5(2):211–215, 1995.

[45] D.G. Brown, R. Visse, G. Sandhu, A. Davies, P.J. Rizkallah, C. Melitz, W.C. Summers, and M.R. Sanderson. Crystal structures of the thymidine kinase from herpes simplex virus type I in complex with deoxythymidine and ganciclovir. *Nature: Structural Biology*, 2(10):876–881, 1995.

[46] S.H. Bryant. Evaluation of threading specificity and accuracy. *PROTEINS: Structure, Function and Genetics*, 26:172–185, 1996.

[47] S.H. Bryant and S.F. Altschul. Statistics of sequence–structure threading. *Current Opinion in Structural Biology*, 5:236–244, 1995.

[48] S.H. Bryant and C.E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *PROTEINS: Structure, Function and Genetics*, 16:92–112, 1993.

[49] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *PROTEINS: Structure, Function and Genetics*, 21:167–195, 1995.

[50] C.J. Bult *et al.* Complete genome sequence of the methanogenic archaeon, *methanococcus jannaschii*. *Science*, 273:1058–1073, 1996.

[51] M. Bycroft, T.J. Hubbard, M. Proctor, S.M. Freund, and A.G. Murzin. The solution structure of the s1 rna binding domain: a member of an ancient nucleic acid-binding fold. *Cell*, 88(2):235–242, 1997.

[52] G. Casari and M.J. Sippl. Structure-derived hydrophobic potential. Hydrophobic potential derived from x-ray structures of globular proteins is able to identify native folds. *Journal of Molecular Biology*, 224:725–732, 1992.

[53] H.S. Chan and K.A. Dill. The effects of internal constraints on the configurations of chain molecules. *Journal of Chemical Physics*, 92(5):3118–3135, 1990.

[54] J.-M. Chandonia and M. Karplus. Neural networks for secondary structure and structural class predictions. *Protein Science*, 4:275–285, 1995.

[55] J.-M. Chandonia and M. Karplus. The importance of larger data sets for protein secondary structure prediction with neural networks. *Protein Science*, 5:768–774, 1996.

[56] K.-M. Chao, R.C. Hardison, and W. Miller. Constrained sequence alignment. *BMB*, 55(3):503–524, 1993.

[57] W. Chiu and M.F. Schmid. Pushing back the limits of electron cryomicroscopy. *Nature: Structural Biology*, 4(5):331–333, 1997.

[58] C. Chothia. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[59] C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5:823–826, 1986.

[60] P.Y. Chou and G.D. Fasman. Conformational parameters for amino acids in helical. β–sheets and random coil regions calculated from proteins. *Biochemistry*, 13:211–222, 1974.

[61] P.Y. Chou and G.D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology*, 47:45–148, 1978.

[62] S.Y. Chung and S. Subbiah. A structural explanation for the twilight zone of protein sequence homology. *Current Biology*, 4:1123–1127, 1996.

[63] R.A. Clayton, O. White, K.A. Ketchum, and J.C. Venter. The first genome from the third domain of life. *Nature*, 387:459–462, 1997.

[64] J. Cohen. The genomics gamble. *Science*, 275:767–782, 1997.

[65] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat, and J.-P. Mornon. Comparison of three algorithms for assignment of secondary structure in proteins: the advantage of a consensu assignment. *Protein Engineering*, 6(4):377–382, 1993.

[66] M.L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16:548–558, 1983.

[67] L.L. Conte and T. Smith. Visible volume: A robust measure for protein structure comparison. *Journal of Molecular Biology*, 273:338–348, 1997.

[68] T.H. Cormen, C.E. Leiserson, and R.L. Rivest. *Introduction to Algorithms*. McGraw-Hill, New York, 1990.

[69] D.G. Covell. Lattice model simulations of polypeptide chain folding. *Journal of Molecular Biology*, 235:1032–1043, 1994.

[70] T.E. Creighton. *Proteins: Structures and Molecular Principles*. Freeman, New York, 1984.

[71] F.H.C. Crick, L. Barnett, S. Brenner, and J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192(4809):1227–1232, 1961.

[72] S. Dalal, S. Balasubramanian, and L. Regan. Protein alchemy: Changing β–sheet into α–helix. *Nature: Structural Biology*, 4(7):548–552, 1997.

[73] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5(Supplement 3):345–352, 1978.

[74] A.F.P. de Araujo and T.C. Pochapsky. Estimates for the potential accuracy required in realistic protein folding simulations and structure recognition experiments. *Folding & Design*, 2:135–139, 1997.

[75] S.E. DeBolt and J. Skolnick. Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non–bonded interactions. *Protein Engineering*, 9(8):637–655, 1996.

[76] D.J. DeRosier. Turn–of–the–century electron microscopy. *Current Biology*, 3:690–692, 1993.

[77] R. S. DeWitte and E. I. Shakhnovich. Pseudodihedrals: Simplified protein backbone representation with knowledge–based energy. *Protein Science*, 3(9):1570–1581, 1994.

[78] V. Di Francesco, J. Garnier, and PJ. Munson. Improving protein secondary structure prediction with aligned homologous sequences. *Protein Science*, 5:106–113, 1996.

[79] K. Diederichs and G.E. Schulz. The three–dimensional structure of the complex between beef heart mitochondrial matrix adenylate kinase and its substrate AMP at 1.85 angstroms resolution. *Journal of Molecular Biology*, 217:541–549, 1991.

[80] K.A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24:1501–1509, 1985.

[81] K.A. Dill and H.S. Chan. From Levinthal to pathways to funnels. *Nature: Structural Biology*, 4(1):10–19, 1997.

[82] L.E. Donate, S.D. Rufino, L.H. Canard, and T.L. Blundell. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Science*, 5(12):2600–2616, 1996.

[83] D. Donnelly, J.P. Overington, and T.L. Blundell. The prediction and orientation of α–helices from sequence alignments: the combined use of environment–dependent substitution tables, fourier transform methods and helix capping rules. *Protein Engineering*, 7(5):645–653, 1994.

[84] R.F. Doolittle. Similar amino acid sequences: Chance or common ancestry? *Science*, 214:149–159, 1981.

[85] D. Dreusicke, P.A. Karplus, and G.E. Schulz. Refined structure of porcine cytosolic adenylate kinase at 2.1 a resolution. *Journal of Molecular Biology*, 199:359–371, 1988.

[86] S.R. Eddy. Hidden markov models. *Current Opinion in Structural Biology*, 6:361–365, 1996.

[87] H. Edelsbrunner. *Algorithms in Combinatorial Geometry*. Springer Verlag, 1987.

[88] H. Edelsbrunner and R. Seidel. Voronoi diagrams and arrangements. *Discrete & Computational Geometry*, 1:25–44, 1986.

[89] A.V. Efimov. Structural trees for protein superfamilies. *PROTEINS: Structure, Function and Genetics*, 28:241–260, 1997.

[90] D. Eisenberg. Into the black of night. *Nature: Structural Biology*, 4(2):95–97, 1997.

[91] D. Eisenberg and A.D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

[92] R.J. Ellis and F.U. Hartl. Protein folding in the cell: Competing models of chaperonine function. *FASEB Journal*, 10:20–26, 1996.

[93] A. Elofsson, D. Fischer, D.W. Rice, S.M. Le Grand, and D. Eisenberg. A study of combined structure/sequence profiles. *Folding & Design*, 1:451–461, 1996.

[94] G.D. Fasman, editor. *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York and London, 1989.

[95] T. Fechteler, U. Dengler, and D. Schomburg. Prediction of protein three–dimensional structures in insertion and deletion regions: A procedure for searching data bases of representative protein fragments using geometric scoring criteria. *Journal of Molecular Biology*, 253:114–131, 1995.

[96] D.R. Ferro and J. Hermans. A different best rigid-body molecular fit routine. *Acta Crystallographica*, A33:345–347, 1977.

[97] K.M. Fiebig and K.A. Dill. Protein core assembly processes. *Journal of Chemical Physics*, 98(4):3475–3478, 1993.

[98] A.V. Finkelstein, A.Y. Badretdinov, and A.M. Gutin. Why do protein architectures have boltzmann-like statistics? *PROTEINS: Structure, Function and Genetics*, 23:142–150, 1995.

[99] A.V. Finkelstein and O.B. Ptitsyn. Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. molec. Biol.*, 50:171–190, 1987.

[100] A.V. Finkelstein and B.A. Reva. A search for the most stable folds in protein chains. *Nature*, 351:497–499, 1991.

[101] D. Fischer, C.-J. Tsai, R. Nussinov, and H. Wolfson. A 3D sequence-independent representation of the protein data bank. *Protein Engineering*, 8:981–997, 1995.

[102] W.M. Fitch and T.F. Smith. Optimal sequence alignments. *Proceedings of the National Academy of Sciences USA*, 80:1382–1386, 1983.

[103] C.M. Fraser *et al.* Genomic seqeunce of a Lyme disease spirochaete, *borrelia burgdorferi*. *Nature*, 390:580–586, 1997.

[104] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *PROTEINS: Structure, Function and Genetics*, 23(4):566–579, 1995.

[105] D. Frishman and P. Argos. Incorporation of non–local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering*, 9(2):133–142, 1996.

[106] D. Frishman and P. Argos. Seventy-five percent accuracy in protein secondary structure prediction. *PROTEINS: Structure, Function and Genetics*, 27:329–335, 1997.

[107] Z. Galil and R. Giancarlo. Speeding up dynamic programming with applications to molecular biology. *Theoretical Computer Science*, 64:107–118, 1989.

[108] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, San Francisco, CA, 1979.

[109] J. Garnier, D.J. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120(1):97–120, 1978.

[110] J. Garnier and B. Robson. Classes of protein folding and accuracy of prediction. In *Workshop on Protein Structure, CECAM, Orsay, France*, pages 147–148, 1979.

[111] D.G. George, W.C. Barker, H.W. Mewes, F. Pfeiffer, and A. Tsugita. The PIR–international protein sequence database. *Nucleic Acid Research*, 24(1):17–20, 1996.

[112] G. Geourjon and G. Deléage. SOPM: a self–optimized method for protein secondary structure prediction. *Protein Engineering*, 7(2):157–164, 1994.

[113] G. Geourjon and G. Deléage. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Computer Applications in Biological Sciences*, 11(6):681–684, 1995.

[114] M. Gerstein, J. Tsai, and M. Levitt. The volume of atoms on the protein surface: Calculated from simulation, using voronoi polyhedra. *Journal of Molecular Biology*, 249:955–966, 1995.

[115] J.-F. Gibrat, T. Madej, and S.H. Bryant. Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, 6:377–385, 1996.

[116] U. Göbel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *PROTEINS: Structure, Function and Genetics*, 18:309 – 317, 1994.

[117] A. Godzik. The structural alignment between two proteins: Is there a unique answer? *Protein Science*, 5:1325–1338, 1996.

[118] A. Godzik, A. Kolinski, and J. Skolnick. Topology fingerprint approach to the inverse protein folding problem. *Journal of Molecular Biology*, 227(1):227–238, 1992.

[119] A. Godzik and J. Skolnick. Sequence-structure matching in globular proteins: applications to supersecondary and tertiary structure determination. *Proceedings of the National Academy of Sciences USA*, 89:12098–12102, 1992.

[120] A. Goffeau. Molecular fish on chips. *Nature*, 385:202–203, 1997.

[121] A. Goffeau *et al.* The yeast genome directory. *Nature*, 387 (suppl.):1–105, 1997.

[122] N. Goldman, J.L. Thorne, and D.T. Jones. Using evolutionary trees in protein secondary structure predictions and other comparative sequence analyses. *Journal of Molecular Biology*, 263:196–208, 1996.

[123] G.H. Gonnet, M.A. Cohen, and S.A. Benner. Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445, 1993.

[124] O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.

[125] O. Gotoh. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinenment as assessed by reference to structural alignments. *Journal of Molecular Biology*, 264:823–838, 1996.

[126] P.A. Greenidge, A. Merz, and G. Folkers. A pseudoreceptor modelling study of the varicella-zoster virus and human thymidine kinase binding sites. *Journal of Computer-Aided Molecular Design*, 9:473–478, 1995.

[127] M. Gribskov, M. Homyak, J. Edenfield, and D. Eisenberg. Profile scanning for three-dimensional structural patterns in protein sequences. *Computer Applications in Biological Sciences*, 4(1):61–66, 1988.

[128] M. Gribskov, A.D. McLachlan, and Eisenberg D. Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences USA*, 84(13):4355–4358, 1987.

[129] H.M. Grindley, P.J. Artymiuk, D.W. Rice, and P. Willet. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology*, 229:707–721, 1993.

[130] M.S. Guyer and F.S. Collins. How is the human genome project doing, and what have we learned so far. *Proceedings of the National Academy of Sciences USA*, 92(November):10841–10848, 1995.

[131] J. Halfmann. Heuristische Verfahren zur Optimierung von Sequenz–Struktur–Alignments von Proteinen mittels empirischer Kontaktpotentiale. Master's thesis, Rheinische Friedrich–Wilhelms–Universität Bonn, 1997.

[132] T.A. Halgren. Potential energy functions. *Current Opinion in Structural Biology*, 5:205–210, 1995.

[133] Y. Harpaz, M. Gerstein, and C. Chothia. Volume changes on protein folding. *Structure*, 2:641–649, 1994.

[134] R.W. Harrison, D. Chatterjee, and I.T. Weber. Analysis of six protein structures predicted by comparative modeling techniques. *PROTEINS: Structure, Function and Genetics*, 23(4):536–547, 1995.

[135] W.E. Hart and S. Istrail. Robust proofs of NP–hardness for protein folding: General lattices and energy potentials. *Journal of Computational Biology*, 4(1):1–22, 1997.

[136] W.E. Hart and S.C. Istrail. Fast protein folding in the hydrophobic–hydrophilic model within three–eights of optimal. *Journal of Computational Biology*, 3(1):53–96, 1996.

[137] W.E. Hart and S.C. Istrail. Lattice and off–lattice side chain models of protein folding: Linear time structure prediction better than 86 % of optimal. Extended Abstract, 1996.

[138] F.U. Hartl and J. Martin. Molecular chaperones in cellular protein folding. *Current Opinion in Structural Biology*, 5:92–102, 1995.

[139] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. Sippl. Identification of native protein folds amongst a large number of incorrect models: The calculation of low energy conformations from potentials of mean force. *Journal of Molecular Biology*, 216:167–180, 1990.

[140] J.P. Hendrik and F.U. Hartl. The role of molecular chaperones in protein folding. *FASEB Journal*, 9:1559–1569, 1995.

[141] J.G. Henikoff and S. Henikoff. Using substitution probabilities to improve position-specific scoring matrices. *Computer Applications in Biological Sciences*, 12(2):135–143, 1996.

[142] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA*, 22(89):10915–10919, 1992.

[143] S. Henikoff and J.G. Henikoff. Automated assembly of protein blocks for database searching. *Nucleic Acid Research*, 19(23):6565–6572, 1993.

[144] S. Henikoff and J.G. Henikoff. Performance evaluation of amino acid substitution matrices. *PROTEINS: Structure, Function and Genetics*, 1(17):49–61, 1993.

[145] J. Heringa, P. Argos, M.R. Egmond, and J. de Vlieg. Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Engineering*, 8(1):21–30, 1995.

[146] D.A. Hinds and M. Levitt. A lattice model for protein structure prediction at low resolution. *Proceedings of the National Academy of Sciences USA*, 89:2536–2540, 1992.

[147] D.S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Communications of the Association for Computing Machinery*, 18(6):341–343, 1975.

[148] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3:522–524, 1994.

[149] L. Holm and C. Sander. Globin fold in a bacterial toxin. *Nature*, 361:309, 1993.

[150] L. Holm and C. Sander. Dali: A network tool for protein structure comparison. *TIBS*, 20:478–480, 1995.

[151] L. Holm and C. Sander. Enzyme HIT. *TIBS*, 22:116–117, 1997.

[152] L. Holm and C. Sander. An evolutionary treasure: Unification of a broad set of amidohydrolases related to urease. *PROTEINS: Structure, Function and Genetics*, 28:72–82, 1997.

[153] B. Honig and F.E. Cohen. Adding backbone to protein folding: why proteins are polypeptides. *Folding & Design*, 1:R17–R20, 1996.

[154] E.S. Huang, S. Subbiah, and M. Levitt. Recognizing native folds by the arrangement of hydrophobic and polar residues. *Journal of Molecular Biology*, 252:709–720, 1995.

[155] E.S. Huang, S. Subbiah, J. Tsai, and M. Levitt. Using a hydrophobic contact potential to evaluate native and near–native folds generated by molecular dynamics simulations. *Journal of Molecular Biology*, 257:716–725, 1996.

[156] T.J.P. Hubbard. Use of β-stand interaction pseudo-potentials in protein structure prediction and modelling. In R.H. Lathrop, editor, *Proceedings of the Biotechnology Computing Track*. IEEE Computer Society Press, 1994.

[157] R. Hughey and A. Krogh. Hidden markov models for sequence analysis: extension and analysis of the basic method. *Computer Applications in Biological Sciences*, 12(2):95–107, 1996.

[158] T. Hunkapiller, R.J. Kaiser, B.F. Koop, and L. Hood. Large-scale and automated dna sequence determination. *Science*, pages 59–67, 1991.

[159] E.G. Hutchinson and J.M. Thornton. A revised set of potentials for β–turn formation in proteins. *Protein Science*, 3:2207–2216, 1994.

[160] S.A. Islam, J. Luo, and M.J.E. Sternberg. Identification and analysis of domains in proteins. *Protein Engineering*, 8(6):523–525, 1995.

[161] Y. Iwata, A. Kasuya, and S. Miyamoto. Reconstruction of 3d coordinates of α–carbon atoms of proteins from a pair of stereographic figures. *Journal of Computer-Aided Molecular Design*, 10:558–566, 1996.

[162] J. Janin. Ångström and calories. *Nature: Structural Biology*, 5:473–479, 1997.

[163] N. Jardine and R. Sibson. *Mathematical Taxonomy*. Wiley and Sons, New York, 1971.

[164] M.S. Johnson and J.P. Overington. A structural basis for sequence comparisons: An evaluation of scoring methodologies. *Journal of Molecular Biology*, 233(4):716–738, 1993.

[165] M.S. Johnson, N. Srinivasan, R. Sowdhamini, and T.L. Blundell. Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.*, 29:1–68, 1994.

[166] D.T. Jones. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Science*, 3:567–574, 1994.

[167] D.T. Jones. `THREADER 2`. User Guide, 1996.

[168] D.T. Jones, W.R. Taylor, and J.M. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[169] D.T. Jones, W.R. Taylor, and J.M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in Biological Sciences*, 8(3):275–282, 1992.

[170] S. Jones, M. Stewart, A. Michie, M.B. Swindells, C. Orengo, and J.M. Thornton. Domain assignment for protein structures using a consensus approach: Characterization and analysis. *Protein Science*, 7:233–242, 1998.

[171] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A32:922–923, 1976.

[172] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A34:827–828, 1978.

[173] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[174] M. Kanehisa. Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan*, 59:34–38, 1996.

[175] S. Karlin and S.F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences USA*, 87(3):2264–2268, 1990.

[176] M. Karplus and J.A. McCammon. The dynamics of proteins. *Scientific American*, 254(4):30–39, 1986.

[177] M. Karplus and A. Šali. Theoretical studies of protein folding and unfolding. *Current Opinion in Structural Biology*, 4(5):58–73, 1995.

[178] L. Kaufmann and P. J. Rousseeuw. *Finding groups in data - An introduction to cluster analysis*. Wiley, 1990.

[179] R.D. King and M.J.E. Sternberg. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science*, 5:2298–2310, 1996.

[180] H.-P. Klenk *et al.* The complete genome seqeunce of a hyperthermophilic, sulphate-reducing archaeon *archaeoglobus fulgidus*. *Nature*, 390:364–370, 1997.

[181] D.K. Klimov and D. Thirumalai. Factors governing the foldability of proteins. *PROTEINS: Structure, Function and Genetics*, 26:411–441, 1996.

[182] N. Kobayashi and N. Go. ATP binding proteins with different folds share a common ATP–binding structural motif. *Nature: Structural Biology*, 4(1):6–7, 1997.

[183] I. Koch, T. Lengauer, and E. Wanke. An algorithm for finding maximal common subtopologies in a set of protein structures. *Journal of Computational Biology*, 3(2):289–306, 1996.

[184] J.-P.A. Kocher, M.J. Rooman, and S.J. Wodak. Factors influencing the ability of knowledge–based potentials to identify native sequence–structure matches. *Journal of Molecular Biology*, 235:1598–1613, 1994.

[185] G. Kolata. Trying to crack the second half of the genetic code. *Science*, 233:1037–1039, 1986.

[186] A. Kolinski and J. Skolnick. Monte carlo simulations of protein folding. I. lattice model and interaction scheme. *PROTEINS: Structure, Function and Genetics*, 18:338–352, 1994.

[187] A. Kolinski and J. Skolnick. Monte carlo simulations of protein folding. II. application to protein a, rop, and crambin. *PROTEINS: Structure, Function and Genetics*, 18:338–352, 1994.

[188] P. Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews*, 93:2395–2417, 1993.

[189] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531, 1994.

[190] F. Kunst *et al.* The complete genome seqeunce of the Gram-positive bacterium *bacillus subtilis*. *Nature*, 390:249–256, 1997.

[191] G. Kuschinsky, H. Lüllmann, and K. Mohr. *Kurzes Lehrbuch der Pharmakologie und Toxikologie*. Georg Thieme Verlag, Stuttgart, New York, 1993.

[192] R.A. Laskowski, M.W. MacArthur, D.S. Moss, and J.M. Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26:283–291, 1993.

[193] R.H. Lathrop. The protein threading problem with sequence amino acid interaction preferences is NP–complete. *Protein Engineering*, 7(9):1059–1068, 1994.

[194] R.H. Lathrop and T.F. Smith. A branch-and-bound algorithm for optimal protein threading with pairwise (contact potential) amino acid interactions. In *Proc. 27th Hawaii Intl Conf. on System Sciences*, pages 1–10, Los Alamitos, CA, 1994. IEEE.

[195] R.H. Lathrop and T.F. Smith. Global optimum protein threading with gapped alignment and empirical pair score functions. *Journal of Molecular Biology*, 255:641–665, 1996.

[196] K.F. Lau and K.A. Dill. Theory of protein mutability and biogenesis. *Proceedings of the National Academy of Sciences USA*, 87:638–642, 1990.

[197] C.M.-R. Lemer, M.J. Rooman, and S.J. Wodak. Protein structure prediction by threading methods: Evaluation of current techniques. *PROTEINS: Structure, Function and Genetics*, 23:337–355, 1995.

[198] T. Lengauer. *Combinatorical Algorithms for Integrated Circuit Layout*. Verlag B. G. Teubner, Stuttgart, 1990.

[199] T. Lengauer. Algorithmic problems in molecular bioinformatics. In *Proceedings of the 2nd Israel Symposium on Theory of Computing and Systems*, pages 177–192. IEEE, 1993.

[200] T. Lengauer, H.T. Mevissen, J. Selbig, R. Thiele, and R. Zimmer. Abschlußbericht der GMD Gruppe des Verbundprojektes „Proteine: Sequenz, Struktur, Evolution (PROTAL)“. URL: cartan.gmd.de/PROTAL/Protal-award/Protal-award.html, 1997.

[201] T. Lengauer, R. Thiele, and R. Zimmer. Modellierung von Proteinstrukturen. *GMD Spiegel*, 1996(2-3):14–18, 1996.

[202] T. Lengauer *et al.* PROTAL. Proteine: Sequenz, Struktur, Evolution. In Wolf, Schmidt, and van der Meer, editors, *Bioinformatik BMBF Statusseminar*, chapter 1, pages 3–26. BMBF, 1995.

[203] C. Levinthal. How proteins fold graciously. In P. Debrunner, J.C.M. Tsibris, and E. Münck, editors, *Mossbauer Spectroscopy in Biological Systems*, pages 22–24. University of Illinois Press, 1969.

[204] M. Levitt. Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology*, 226(2):507–533, 1992.

[205] M. Levitt. Competitive assessment of protein fold recognition and alignment accuracy. *PROTEINS: Structure, Function and Genetics*, pages 92–104, 1997.

[206] M. Levitt and J. Greer. Automatic identification of secondary structure in globular proteins. *Journal of Molecular Biology*, 114(2):181–239, 1977.

[207] D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985.

[208] Y. Luo, L. Lai, X. Xu, and Y. Tang. Defining topological equivalences in protein structures by means of dynamic programming algorithm. *Protein Engineering*, 6(4):373–376, 1993.

[209] R. Lüthy, J.U. Bowie, and D. Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356:83–85, 1992.

[210] R. Lüthy, A.D. McLachlan, and D. Eisenberg. Secondary structure–based profiles: Use of structure–conserving scoring tables in searching protein sequence databases for structural similarities. *PROTEINS: Structure, Function and Genetics*, 10:229–239, 1991.

[211] T. Madej, M.S. Boguski, and S.H. Bryant. Threading analysis suggests that the obese gene product may be a helical cytokine. *FEBS Letters*, 373:13–18, 1995.

[212] V.N. Maiorov and G.M. Crippen. Contact potential that recognizes the correct folding of globular proteins. *Journal of Molecular Biology*, 227:876–888, 1992.

[213] E.J. Mancini, F. de Haas, and S.D. Fuller. High-resolution icosahedral reconstruction: fulfilling the promise of cryo-electron microscopy. *Structure*, 5:741–750, 1997.

[214] M.L. Mansfield. Are there knots in proteins? *Nature: Structural Biology*, 1(4):213–214, 1994.

[215] M.L. Mansfield. Fit to be tied. *Nature: Structural Biology*, 4(3):166–167, 1997.

[216] A. Marchler-Bauer and S.H. Bryant. A measure of success in fold recognition. *Trends in Biochemical Sciences*, 22:236–240, 1997.

[217] A. Marchler-Bauer, M. Levitt, and S.H. Bryant. A retrospective analysis of CASP2 threading predictions. *PROTEINS: Structure, Function and Genetics*, Supplement 1:83–91, 1997.

[218] A.E. Mark and W.F. van Gunsteren. Decomposition of the free energy of a system in terms of specific interactions: Implications for theoretical and experimental studies. *Journal of Molecular Biology*, 240:167–176, 1994.

[219] A.C.R. Martin and J.M. Thornton. Structural families in loops of homologous proteins: Automatic classification, modelling and application to antibodies. *Journal of Molecular Biology*, 263:800–815, 1996.

[220] J. Martin and F.U. Hartl. Chaperone–assisted protein folding. *Current Opinion in Structural Biology*, 7:41–52, 1997.

[221] Y. Matsuo and K. Nishikawa. Protein structural similarities predicted by a sequence-structure compatibility method. *Protein Science*, 3:2055–2063, 1994.

[222] A.C.W. May. Pairwise iterative superposition of distantly related proteins and assessment of the significance of 3–D structural similarity. *Protein Engineering*, 9(12):1093–1101, 1996.

[223] A.C.W. May and M.S. Johnson. Improved genetic algorithm–based protein structure comparisons: pairwise and multiple superpositions. *Protein Engineering*, 8(9):873–882, 1995.

[224] A.D. McLachlan. Rapid comparison of protein structures. *Acta Crystallographica*, A 38:871–873, 1982.

[225] F. Melo and E. Feytmans. Novel knowledge–based mean force potential at atomic level. *Journal of Molecular Biology*, 267:207–222, 1997.

[226] H. Mevissen, R. Thiele, R. Zimmer, and T. Lengauer. The ToPLign software environment – Toolbox for protein alignment. In *Bioinformatik '94*. Jena, IMB – Institut für molekulare Biotechnologie, 1994.

[227] H.T. Mevissen and M. Vingron. Quantifying the local reliability of a sequence alignment. *Protein Engineering*, 9(2):127–132, 1996.

[228] A.D. Michie, C.A. Orengo, and J.M. Thornton. Analysis of domain structural class using an automated class assignment protocol. *Journal of Molecular Biology*, 262:168–185, 1996.

[229] W. Miller and E.W. Myers. Sequence comparison with concave weighting functions. *BMB*, 50:97–120, 1988.

[230] A.D. Miranker and C.M. Dobson. Collapse and cooperativity in protein folding. *Current Opinion in Structural Biology*, 6:31–42, 1995.

[231] L.A. Mirny, V. Abkevich, and E.I. Shakhnovich. Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model. *Folding & Design*, 1(2):103–116, 1996.

[232] S. Miyazawa and R.L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structure: Quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[233] S. Miyazawa and R.L. Jernigan. Residue-residue potentials with favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256:623–644, 1996.

[234] M. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6:525–533, 1993.

[235] B. Morgenstern, A. Dress, and T. Werner. Multiple DNA and protein sequence alignment based on segment–to–segment comparison. *Proceedings of the National Academy of Sciences USA*, 93(22):12098–12103, 1996.

[236] D.R Morrison. PATRICIA–practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM*, pages 514–534, 1968.

[237] S. Mosimann, R. Meleshko, and M.N.G. James. A critical assessment of comparative molecular modelling of tertiary structures of proteins. *PROTEINS: Structure, Function and Genetics*, 23(3):301–317, 1995.

[238] J. Moult. Comparison of database potentials and molecular mechanics force fields. *Current Opinion in Structural Biology*, 25:194–199, 1996.

[239] J. Moult, T. Hubbard, S.H. Bryant, K. Fidelis, and J.T. Pederson. Critical assessment of methods of protein structure prediction (CASP): Round ii. *PROTEINS: Structure, Function and Genetics*, Supplement 1:2–6, 1997.

[240] J. Moult, J.T. Pederson, R. Judson, and K. Fidelis. A large–scale experiment to assess protein structure prediction methods. *PROTEINS: Structure, Function and Genetics*, 23:i–iv, 1995.

[241] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. Specific enzymatic amplification of dna in vitro: the polymerase chain reaction. In *Cold Spring Harbour Symposium Quant Biology*, volume 51, Pt 1, pages 263–273, 1986.

[242] A.G. Murzin. Structural classification of proteins: new superfamilies. *Current Opinion in Structural Biology*, 6:386–394, 1996.

[243] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[244] E. W. Myers and W. Miller. Approximate matching of regular expressions. *BMB*, 51:5–37, 1989.

[245] E.W. Myers and W. Miller. Optimal alignments in linear space. *Computer Applications in Biological Sciences*, 4(1):11–17, 1988.

[246] G. Myers, S. Selznick, Z. Zhang, and W. Miller. Progressive multiple alignment with constraints. *Journal of Computational Biology*, 3(4):563–572, 1996.

[247] D. Naor and D. Brutlag. On suboptimal alignments of biological sequences. In *Proceedings of the 4th Symposium on Combinatorial Pattern Matching*. Lecture Notes in Computer Science, 1993.

[248] D. Naor and D.L. Brutlag. On near-optimal alignments of biological sequences. *Journal of Computational Biology*, 1(4):349–366, 1994.

[249] D. Naor, D. Fischer, R.L. Jergnigan, H.J. Wolfson, and R. Nussinov. Amino acid pair interchanges at spatially conserved locations. *Journal of Molecular Biology*, 256(4):924–938, 1996.

[250] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[251] E. Neher. How frequent are correlated changes in families of protein sequences. *Proceedings of the National Academy of Sciences USA*, 91:98–102, 1994.

[252] W.J. Netzer and F.U. Hartl. Recombination of protein domains facilitated by co–translational folding in eukaryotes. *Nature*, 388:343–349, 1997.

[253] K. Niefind and D. Schomburg. Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. *Journal of Molecular Biology*, 219:481–497, 1991.

[254] M.W. Nirenberg, P Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, and C. O'Neal. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences USA*, 53:1161–1168, 1965.

[255] M.W. Nirenberg and J.H. Matthaei. The dependence of cell–free protein synthesis in *e. coli* upon naturally occuring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences USA*, 47:1588–1602, 1961.

[256] S. Nishimura, D.S. Jones, and H.G. Khorana. Studies on polynucleotides. 48. the in vitro synthesis of a co–polypeptide containing two amino acids in alternating sequence dependent upon a DNA–like polymer containing two nucleotides in alternating sequence. *Journal of Molecular Biology*, 13:302–324, 1965.

[257] R. Nussinov and H.J. Wolfson. Efficient detection of three–dimensional structural motifs in biological macromolecules by computer vision techniques. *Proceedings of the National Academy of Sciences USA*, 88:10495–10499, 1991.

[258] B. Oliva, P.A. Bates, E. Querol, F.X. Avilés, and M.J. Sternberg. An automated classification of the structure of protein loops. *Journal of Molecular Biology*, 266(4):814–830, 1997.

[259] O. Olmea and A. Valencia. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & Design*, 2:S25–S32, 1997.

[260] C.A. Orengo, T.P. Flores, W.R. Taylor, and J.M. Thornton. Identification and classification of protein fold families. *Protein Engineering*, 6(5):485–500, 1993.

[261] C.A. Orengo, D.T. Jones, and J.M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372:631–634, 1994.

[262] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. Cath – a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[263] C.A. Orengo, Brown N.P., and W.R. Taylor. Fast structure alignment for protein databank searching. *PROTEINS: Structure, Function and Genetics*, 14(2):139–167, 1992.

[264] C.A. Orengo and W. Taylor. A rapid method of protein structure alignment. *Journal of theoretical Biology*, 147:517–551, 1990.

[265] J. O'Rourke. *Computational Geometry in C.* Cambridge University Press, 1994.

[266] C. Ouzounis, C. Sander, M. Scharf, and R. Schneider. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3d structures. *Journal of Molecular Biology*, 232:000–021, 1993.

[267] J.P. Overington, D. Donnelly, M.S. Johnson, A. Šali, and Tom L. Blundell. Environment–specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *PS*, 1:216–226, 1992.

[268] B. Park, E.S. Huang, and M. Levitt. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *Journal of Molecular Biology*, 266:831–846, 1997.

[269] B. Park and M. Levitt. Energy functions that discriminate x–ray and near–native folds from well–constructed decoys. *Journal of Molecular Biology*, 258:367–392, 1996.

[270] S. Pascarella and P. Argos. A data bank merging related protein structures and sequences. *Protein Engineering*, 5(2):121–137, 1992.

[271] L. Pauling. *The Nature of Chemical Bond.* Cornell University Press, Ithaca, NY, 1939.

[272] K. Pawlowski, A. Bierzynski, and A. Gozik. Structural diversity in a family of homologous proteins. *Journal of Molecular Biology*, 258:349–366, 1996.

[273] W.R. Pearson. Comparison of methods for searching protein sequence databases. *Protein Science*, 4(6):1145–1160, 1995.

[274] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences USA*, 85(8):2444–2448, 1988.

[275] E. Pebay-Peyroula, G. Rummel, J.P. Rosenbusch, and E.M Landau. X–ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science*, 277:1676–1681, 1997.

[276] J.T. Pederson and J. Moult. Protein folding simulations with genetic algorithms and a detailed molecular description. *Journal of Molecular Biology*, 269:240–259, 1997.

[277] J. Pontius, J. Richelle, and S.J. Wodak. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of Molecular Biology*, 264:121–136, 1996.

[278] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1993.

[279] N.A. Ranson, N.J. Dunster, S.G. Burston, and A.R. Clarke. Chaperonins can catalyse the reversal of early aggregation steps when a protein misfolds. *Journal of Molecular Biology*, 250:581–586, 1995.

[280] B.A. Reva, M.F. Sanner A.V. Finkelstein, and A.J. Olson. Accurate mean-force pairwise-residue potentials for discrimination of protein folds. In *Pacific Symposium on Biocomputing 1997*, 1997.

[281] B.A. Reva, Alexei V. Finkelstein, and Jeffrey Skolnick. What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? *Folding & Design*, 3(2):141–147, 1998.

[282] B.A. Reva, A.V. Finkelstein, D.S. Rykunov, and A.J. Olson. Building self-avoiding lattice models of proteins using a self-consistent field optimization. *PROTEINS: Structure, Function and Genetics*, 26:1–8, 1996.

[283] B.A. Reva, A.V. Finkelstein, M.F. Sanner, and A.J. Olson. Adjusting potential energy functions for lattice models of chain molecules. *PROTEINS: Structure, Function and Genetics*, 25:379–388, 1996.

[284] D.W. Rice and D. Eisenberg. A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *Journal of Molecular Biology*, 267:1026–1038, 1997.

[285] F.M. Richards and C.E. Kundrot. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *PROTEINS: Structure, Function and Genetics*, 3(2):71–84, 1988.

[286] P.A. Rioux, W.A. Gilbert, and T.G. Littlejohn. A portable search engine and browser for the entrez database. *Journal of Computational Biology*, 1(4):293–295, 1994.

[287] B. Rost, P. Fariselli, and R. Casadio. Topology prediction for helical transmembrane proteins at 86 accuracy. *Protein Science*, 5:1704–1718, 1996.

[288] B. Rost and C. Sander. Improved prediction of protein secondary structure by use of sequennce profiles and neural networks. *Proceedings of the National Academy of Sciences USA*, 90:7558–7562, 1993.

[289] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232:584–599, 1993.

[290] B. Rost and C. Sander. Conservation and prediction of solvent accessibility in protein families. *PROTEINS: Structure, Function and Genetics*, 20:216–226, 1994.

[291] B. Rost, C. Sander, and R. Schneider. Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, 235:13–26, 1994.

[292] B. Rost, R. Schneider, and C. Sander. Protein fold recognition by prediction–based threading. *Journal of Molecular Biology*, 270:471–480, 1997.

[293] D.A. Rozwarski, A.M. Gronenborn, Clore G.M., Bazan J.F., A. Bohm, Wlodawer A., M. Hatada, and P.A. Karplus. Structural comparisons among the short-chain helical cytokines. *Structure*, 2(3):159–173, 1994.

[294] R.W. Ruddon, S.A. Sherman, and E. Bedows. Protein folding in the endoplasmic reticulum: lessons from the human chorionic gonadotropin β subunit. *Protein Science*, 5:1443–1452, 1996.

[295] S.D. Rufino, L.E. Donate, L.H. Canard, and T.L. Blundell. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. *Journal of Molecular Biology*, 267(2):2600–2616, 1997.

[296] D. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. *Parallel distributed processing*, 1:318–362, 1986.

[297] R.B. Russel and M.J.E. Sternberg. Two new examples of protein structural similarities within the structure–function twilight zone. *Protein Engineering*, 10(4):333–338, 1997.

[298] R.B. Russell and G. Barton. Structural features can be unconserved in proteins with similar folds. *Journal of Molecular Biology*, 244(3):332–350, 1994.

[299] R.B. Russell, R.R. Copley, and G.J. Barton. Protein fold recognition by mapping predicted secondary structures. *Journal of Molecular Biology*, 259:349–365, 1996.

[300] D.S. Rykunov, B.A. Reva, and A.V. Finkelstein. Accurate general method for lattice approximation of three-dimensional structure of a chain molecule. *PROTEINS: Structure, Function and Genetics*, 22:100–109, 1995.

[301] A.A. Salamov and V.V. Solovyev. Prediction of protein secondary structure by combining nearest-neighbour algorithms and multiple sequence alignment. *Journal of Molecular Biology*, 247:11–15, 1995.

[302] A.A. Salamov and Solovyev V.V. Protein secondary structure prediction using local alignments. *Journal of Molecular Biology*, 268(1):31–36, 1997.

[303] A. Šali. Modeling mutations and homologous proteins. *Current Opinion in Biotechnology*, 6(4):437–451, 1995.

[304] A. Šali and T.L. Blundell. Definition of general topological equivalence in protein structures. *Journal of Molecular Biology*, 212:403–428, 1990.

[305] A. Šali and T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234:779–815, 1993.

[306] A. Šali and J.P. Overington. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Science*, 3:1582–1596, 1994.

[307] A. Šali, E. Shaknovich, and M. Karplus. How does a protein fold ? *Nature*, 369:248–251, 1994.

[308] A. Šali, E. Shaknovich, and M. Karplus. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *Journal of Molecular Biology*, 235:1614–1636, 1994.

[309] R. Sánchez and A. Šali. Advances in comparative protein–structure modelling. *Current Opinion in Structural Biology*, 7:206–214, 1997.

[310] C. Sander and R. Schneider. Database of homology derived protein structures and the structural meaning of sequence alignment. *PROTEINS: Structure, Function and Genetics*, 9:56–68, 1991.

[311] H. Schindelin, W. Jiang, M. Inouye, and U. Heinemann. Crystal structure of CspA, the major cold shock protein of escherichia coli. *Proceedings of the National Academy of Sciences USA*, 91(11):5119–5123, 1994.

[312] J. Selbig. Contact pattern-induced pair potentials for protein fold recognition. *Protein Engineering*, 8(4):339–351, 1995.

[313] J. Selbig and P. Argos. Relationships between protein sequence and structure patterns based on residue contacts. *PROTEINS: Structure, Function and Genetics*, 31:172–185, 1998.

[314] J. Setubal and J. Meidani, editors. *Introduction to Computational Molecular Biology*. PWS Publishing Company, Boston, 1997.

[315] I.N. Shindyalov, N.A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering*, 7:349–358, 1994.

[316] D. Shortle, Y. Wang, J.R. Gillespie, and J.O. Wrabl. Protein folding for realists: A timeless phenomenon. *Protein Science*, 5:991–1000, 1996.

[317] A.S. Siddiqui and G.J. Barton. Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. *Protein Science*, 4:872–884, 1995.

[318] M. Singer and P. Berg. *Gene und Genome*. Spektrum Akademischer Verlag, 1992.

[319] M.J. Sippl. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, 213:859–883, 1990.

[320] M.J. Sippl. Boltzmann's principle, knowledge–based mean fields and protein folding. An approach to the computational determination of protein structures. *Journal of computer-aided molecular design*, 7:473–501, 1993.

[321] M.J. Sippl. Recognition of errors in three–dimensional structures of proteins. *PROTEINS: Structure, Function and Genetics*, pages 1–8, 1993.

[322] M.J. Sippl. Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, 5:229–235, 1995.

[323] M.J. Sippl. Helmholtz free energy of peptide hydrogen bonds in proteins. *Journal of Molecular Biology*, 260:644–648, 1996.

[324] M.J. Sippl, M. Ortner, M. Jaritz, P. Lachner, and H. Flöckner. Helmholtz free energy of atom pair interactions in proteins. *Folding & Design*, 1:289–298, 1996.

[325] M.J. Sippl and S. Weitckus. Detection of native-like models for amino acid sequence of unknown three-dimensional structure in a data base of known protein conformations. *PROTEINS: Structure, Function and Genetics*, 13:258–271, 1992.

[326] H. Sklenar, C. Etchebest, and R. Lavery. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *PROTEINS: Structure, Function and Genetics*, 6(1):46–60, 1989.

[327] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1980.

[328] R. Sowdhamini, S.D. Rufino, and T.L. Blundell. A database of globular protein structural domains: clustering of representative family members into similar folds. *Folding & Design*, 1:209–220, 1996.

[329] J. Stoye, V. Moulton, and A.W. Dress. DCA: An efficient implementation of the divide–and–conquer approach to simultaneous multiple sequence alignment. *Computer Applications in Biological Sciences*, 13(6):625–626, 1997.

[330] S.S. Sturrock and J.F. Collins. Mpsrch version 1.3. Technical report, Biocomputing Research Unit, University of Edinburgh, UK., 1993.

[331] S. Sudarsanam. Structural diversity of seqeunteially identical subsequences of proteins: Identical octapeptides can have different conformations. *PROTEINS: Structure, Function and Genetics*, 30:228–231, 1998.

[332] M.J. Sutcliffe, I. Haneef, D. Carney, and T.L. Blundell. Knowledge based modelling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engineering*, 1:377–384, 1987.

[333] M. Suyama, Y. Matsuo, and K. Nishikawa. Comparison of protein structures using 3D profile alignment. *Journal of Molecular Evolution*, 44 (Suppl. 1):163–173, 1997.

[334] M.B. Swindells. A procedure for detecting structural domains in proteins. *Protein Science*, 4:103–112, 1995.

[335] M.B. Swindells. A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Science*, 4:93–102, 1995.

[336] M.B. Swindells, M.W. MacArthur, and J.M. Thornton. Intrinsic ϕ, ψ propensities of amino acids, derived from the coil regions of known structures. *Nature: Structural Biology*, 2(7):596–603, 1995.

[337] M. Szardenings, B. Vase, H.-J. Hecht, J. Collins, and D. Schomburg. Highly effective protease inhibitors from variants of human pancreatic secretory trypsin inhibitor (hPSTI): an assessment of 3–d structure-based protein design. *Protein Engineering*, 8(1):45–52, 1995.

[338] W. Taylor and C.A. Orengo. A holistic approach to protein structure alignment. *Protein Engineering*, 2(7):505–519, 1989.

[339] W. Taylor and C.A. Orengo. Protein structure alignment. *Journal of Molecular Biology*, 208:1–22, 1989.

[340] R. Thiele, R. Zimmer, and T. Lengauer. Recursive dynamic programming for adaptive sequence and structure alignment. In C. Rawlings *et al.*, editor, *Intelligent Systems for Molecular Biology*, pages 384–392, Cambridge, UK, 1995. AAAI Press.

[341] J.-F. *et al.* Tomb. The complete genome sequence of the gastric pathogen *helicobacter pylori*. *Nature*, 388:539–547, 1997.

[342] U. Tönges, S.W. Perrey, J. Stoye, and A.W. Dress. A general method for fast multiple sequence alignment. *Gene*, 172(1):GC33–GC41, 1996.

[343] J.P. Turkenburg and E.J. Dodson. Modern developments in molecular replacement. *Current Opinion in Structural Biology*, 6:604–610, 1996.

[344] De Filipis V., C. Sander, and G. Vriend. Predicting local structural changes that result from point mutations. *Protein Engineering*, 7:1203–1208, 1994.

[345] W. F. van Gunsteren and P. K. Weiner, editors. *Computer Simulation of Biomolecular Systems*, volume 1. Escom, Leiden, 1989.

[346] W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, editors. *Computer Simulation of Biomolecular Systems*, volume 2. Escom, Leiden, 1993.

[347] W.F. van Gunsteren and H.J.C. Berendsen. *Groningen Molecular Simulation (GROMOS)* Library Manual. Biomos, Groningen, 1987.

[348] W.F. van Gunsteren and H.J.C. Berendsen. Moleküldynamik-computersimulation: Methodik, Anwendungen und Perspektiven in der Chemie. *Angewandte Chemie*, 102:1020–1055, 1990.

[349] H.W.T. van Vlijmen and M. Karplus. PDB–based protein loop prediction: Parameters for selection and methods for optimization. *Journal of Molecular Biology*, 267:975–1001, 1997.

[350] M. Vásquez. Modeling side–chain conformation. *Current Opinion in Structural Biology*, 6:217–221, 1996.

[351] M. Vingron. Near–optimal sequence alignments. *Current Opinion in Structural Biology*, 6:346–352, 1996.

[352] M. Vingron and M.S. Waterman. Sequences alignments and penalty choice: Review of concepts, case studies and implications. *Journal of Molecular Biology*, 235:1–12, 1994.

[353] D. Voet and J.G. Voet. *Biochemie*. VCH, Weinheim, 1992.

[354] G. Vogt, T. Etzold, and P. Argos. An assessment of amino acid exhange matrices in aligning protein sequences: The twilight zone revisited. *Journal of Molecular Biology*, 249:816–831, 1995.

[355] G.F. Voronoi. Nouveles applications des paramétres continus à la théorie des formes quadratique. *J. Reine Angew. Math.*, 134:198–287, 1908.

[356] H. Wako and T.L. Blundell. Use of amino acid environment–dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins: I. solvent accessibility classes. *Journal of Molecular Biology*, 238(5):682–692, 1994.

[357] H. Wako and T.L. Blundell. Use of amino acid environment–dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins: II. secondary structures. *Journal of Molecular Biology*, 238(5):693–708, 1994.

[358] Z.-X. Wang. How many fold types of protein are there in nature? *PROTEINS: Structure, Function and Genetics*, 26:186–191, 1996.

[359] M.S. Waterman and M. Eggert. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparison. *Journal of Molecular Biology*, 197(4):723–728, 1987.

[360] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Jr. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106:765, 1984.

[361] J. White, I. Muchnik, and T.F. Smith. Modeling protein cores with markov random fields. *Mathematical Biosciences*, 124:149–179, 1994.

[362] R. Wilber. The concave least-weight subsequence problem revisited. *Journal of Algorithms*, 9:418–425, 1988.

[363] K. Wild, T. Bohner, A. Aubry, G. Folkers, and G.E. Schulz. The three-dimensional struture of thymidine kinase from herpes simplex virus type I. *FEBS Letters*, 368:289–292, 1995.

[364] M. Wilm, A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis, and M. Mann. Femtomole sequencing of proteins from polyacrylamide gels by nano–electrospray mass spectrometry. *Nature*, 379:466–469, 1996.

[365] M. Wilmanns and D. Eisenberg. Inverse protein folding by residue pair preference profile method: estimating the correctness of alignments of structurally compatible sequences. *Protein Engineering*, 8(7):627–639, 1995.

[366] M. Wöhler. Optimierung von Alignment–Bewertungssystemen mit Voronoi Zerlegungen von Proteinen. Master's thesis, Rheinische Friedrich–Wilhelms–Universität Bonn, 1997.

[367] M. Wöhler, R. Thiele, and R. Zimmer. New scoring schemes for protein fold recognition based on voronoi contacts. In *German Conference on Bioinformatics (GCB'97)*, 1997.

[368] K. Wüthrich. *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, New York, N.Y., 1986.

[369] Y. Xu and E.C. Uberbacher. A polnomial-time algorithm for a class of protein threading problems. *Computer Applications in Biological Sciences*, 12:511–517, 1996.

[370] F.F. Yao. *Handbook of Theoretical Computer Science. Algorithms and Complexity*, volume A, chapter Computational Geometry, pages 525–631. North-Holland, New York, N. Y., 1990.

[371] T.-M. Yi and E.S. Lander. Recognition of related proteins by iterative template refinement (ITR). *Protein Science*, 3:1315–1328, 1994.

[372] K. Yue and K.A. Dill. Inverse protein folding problem: Designing polymer sequences. *Proceedings of the National Academy of Sciences USA*, 89:4163–4167, 1992.

[373] B. Zhang, L. Jaroszewski, L. Rychlewski, and A. Godzik. Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies. *Folding & Design*, 2(5):307–317, 1997.

[374] K.Y.J. Zhang and D. Eisenberg. The three-dimensional profile method using residue preference as a continuous function of residue environment. *Protein Science*, 3:687–695, 1994.

[375] F. Zhang *et al.* Crystal structure of the *obese* protein leptin-E100. *Nature*, 387:206–209, 1997.

[376] Q. Zheng and D.J. Kyle. Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based on extensive and multiple copy conformational samplings. *PROTEINS: Structure, Function and Genetics*, 24(2):209–217, 1996.

[377] Z.-Y Zhu and T.L. Blundell. The use of amino acid patterns of classified helices and strands in secondary structure predition. *Journal of Molecular Biology*, 260:261–276, 1996.

[378] A. Zien. Optimierungsmethoden zur Kalibrierung empirischer Kostenfunktionen. Master's thesis, Rheinische Friedrich–Wilhelms–Universität Bonn, 1997.

[379] R. Zimmer. Fold recognition by 123D. personal communication, 1997.

[380] R. Zimmer and T. Lengauer. Fast and numerically stable parametric alignment of biosequences. In M. Waterman, editor, *1st Ann. Int. Conf. on Computational Molecular Biology (RECOMB'97)*, pages 344–353. ACM Press, 1997.

[381] R. Zimmer and R. Thiele. Fast protein fold recognition and accurate sequence to structure alignment. In *Bioinformatics: German Conference on Bioinformatics (GCB'96)*, volume 1278, pages 137–146. Lecture Notes in Computer Science, 1997.

[382] R. Zimmer, M. Wöhler, and R. Thiele. New scoring schemes for protein fold recognition based on voronoi contacts. *Bioinformatics*, 14(3):295–308, 1998.

[383] F. Zu-Kang and M.J. Sippl. Optimum superimposition of protein structures: ambiguities and implications. *Folding & Design*, 1:123–132, 1996.