Adaptation of the mining environment SCAIView to the needs of animal scientists

Master Thesis *by*: **Sudeep Sahadevan**
Life Science Informatics Matr. number 2046806
B-IT
Dahlmannstrasse 2
D-53113 Bonn

and done at:

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Schloß Birlinghoven
D-53754 Sankt Augustin

and:

Institute of Animal Science Animal Breeding and Husbandry and Animal Genetics Group
Endenicher Allee 15
D-53115 Bonn

Professor : Prof. Dr. Martin Hofmann–Apitius
University of Bonn, Applied Bioinformatics
Fraunhofer SCAI, Department of Bioinformatics

2nd Professor : Prof. Dr. Karl Schellander
University of Bonn, Agricultural Faculty
Institute of Animal Science
Animal Breeding and Husbandry Group

Supervisor : Dr.-Ing. Christoph M. Friedrich
Fraunhofer SCAI, Department of Bioinformatics

Declaration

I herewith declare that I have prepared the present Master Thesis on my own and without any further sources and aides but those cited.

Sankt Augustin, December 2, 2009

Sudeep Sahadevan

Acknowledgement

First of all, I want to thank Fraunhofer SCAI and Institute of Animal Science Animal Breeding and Husbandry Group / Animal Genetics Group for this collaborated project.

I would like to extend my gratitude towards Prof. Dr. Martin Hofmann-Apitius for giving me a chance to work on this project and also his support throughout the course.

I would also like to thank Prof. Dr. Karl Schellander for being the second professor for this work and also for his help and support during this project.

I am grateful to Dr.-Ing. Christoph M. Friedrich for his support, guidance, discussions and feed backs throughout this entire work.

I am particularly thankful to Dr. Juliane Fluck, Dr. Dawit Tesfaye, Heinz-Theodor Mevissen and Roman Klinger for their help and valuable advice.

I would like to acknowledge Oriol Fornés for his technical support with the BIANA network.

I would also like to thank Fraunhofer SCAI Department of Bioinformatics team and Institute of Animal Science Animal Breeding and Husbandry Group / Animal Genetics Group members for their support.

Finally, I owe my deepest gratitude to my family and friends for help and support throughout this thesis and the entire course.

Abstract

Researches in biological sciences are knowledge based, where any prior knowledge in a given research field is as important as the formulation of a hypothesis. Text mining and knowledge discovery methods have been utilized in human genomics and related fields for identifying genes, proteins, related networks and to enrich the identified entities with external database information, the important prior knowledge in the field. But, in livestock genomics however, the extraction of prior information is restricted to traditional keyword search and document retrieval. The aim of this thesis is the introduction of text mining applications to livestock genomics and associated field. ProMiner, a dictionary and rule based named entity recognition system and the associated knowledge dscovery platform SCAIView has been successfully used in human, mouse and plant genomics and related fields. Through this thesis the existing SCAIView and the associated system was adapted for the use in livestock genomics field and the present version works with cattle and pig data. Considering the importance given to the preimplantation period in cattle genomics, a terminology for cattle preimplantation period was developed and integrated into the adapted livestock genomics version of SCAIView to aid concept based search. The present version of livestock genomics SCAIView is intended as a prototype for demonstrating possibilities of text mining and knowledge discovery to the researchers in livestock genomics.

Contents

List of Figures

List of Tables

1 Introduction

The practice of raising livestock has been in existence from a very early time in human history. Yet how much do we know about them? Most of the livestock animals have been a major source of protein, fiber (wool, hides) and labour since domestication. Besides these qualities, farm animals can also be used as models for pathological and physiological studies, since the physiology and anatomy of farm animals like pig, sheep and even cattle are similar to that of human beings. Also, there are less ethical issues concerned with the use of livestock animals for these studies. The phenotypic diversity and large population of these animals make them suitable candidates for wide scale genome analysis. The study and analysis of genes and proteins of these animals are of a wide interest to researchers, as the knowledge gained from the study of genome and proteome analysis can be used to make these animals more healthy and high yielding. Many researchers are interested in farm animal genomics because of the benefits of understanding proteomics and genomics of various organisms (Roberts et al., 2009). Numerous instances can be pointed out where domesticated farm animals could be effectively used as model organisms. Domesticated sheep species are widely used for studying the effect of environmental factors on developing embryo. Piglets are used as model organisms for human infant nutrition and pigs are also prone to diseases like artherosclerosis, gastric ulcers and obesity like humans, which makes them a model organism for these diseases (Roberts et al., 2009). Hence, it can be said that livestock genomics can play a major role in improving the overall quality of human life.

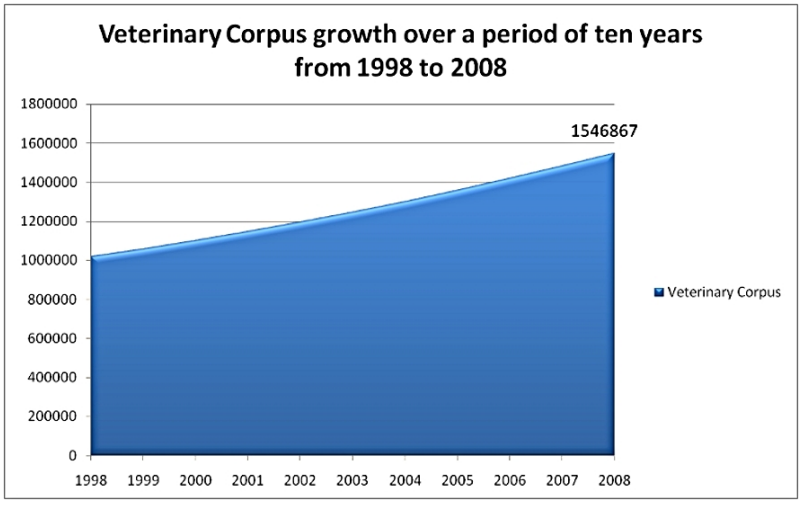


Figure 1.1: Chart showing growth veterinary corpus in PubMed over a period of ten years from 1998 to 2008

In the whole animal genomics and proteomics domain, what is the role of bioinformatics tools and other computational techniques? Bioinformatics tools, computational techniques and other information retrieval systems come into picture in data storage, data retrieval and data analysis from databases. A most common example is PubMed[1] (refer to Section 3.1.2), one of the largest sources of electronic biomedical data. In Pubmed, the number of animal science and animal genomics documents show a steady increase. PubMed has a special collection of documents relating to the animal science domain, the Veterinary corpus[2] and the corpus is presently a collection of more than 1.5 million documents. Figure 1.1 shows the increase in number of documents in veterinary corpus over a period of ten years and Figure 1.3 shows the number of documents present in the corpus for organisms such as cattle, mouse and pig. The use of high throughput technologies in livestock genomics has lead to an ever-increasing amount of data in the livestock genomics field. But, to make use of these additional information, enrichment with external information is necessary. For example, overlaying functional gene annotation data on micro array experimental data. However, most of the "enrichment" information needed is present in free written texts like scientific publications, not in databases as structured information. This information in free text is often considered as hidden information, and the search and retrieval of this information are not trivial
(Leser and Hakenberg, 2005). Text mining is the process of retrieving high quality, "hidden" information from written texts. Knowledge discovery is a term that is used quite often with text mining and is defined as the extraction of previously unknown, hidden and potentially useful information. In a broad sense of view text mining can be categorized as one of the methods for knowledge discovery. Powerful text mining tools are in existence in the human genomics field, which could extract high quality gene and protein information from free texts. A classical example of such a system is ProMiner (Hanisch et al., 2005), a dictionary and rule based text mining tool and associated search engine SCAIView (Hofmann-Apitius et al., 2008), developed at the Fraunhofer Institute of Algorithms and Scientific Computing (SCAI). An example of knowledge discovery through text mining can be demonstrated through the ABC model of complementarity, (Laine, 2008), (Swanson and Smalheiser, 1997) where two concepts from different domains are joined thorough a set of concepts that is common to both the domains. This was explained through demonstrating connection between fish oil and Multiple Sclerosis (MS). A component of fish oil is its omega 3 essential fatty acid- "docosahexaenoic acid", which has beneficial effects on the disease, yet without proofs then. When a PubMed search was conducted with "Multiple Sclerosis" and, "docosahexaenoic acid" as individual terms, the search returned 38113 for the former and 4370 for the latter. But for the query "Multiple Sclerosis AND docosahexaenoic acid" the query returned 8 articles as hits. But a close inspection of the contents of the initial search results reveals that the search terms has a set of concepts and terms in common. For example, the biological entities like "MMP9" and "TIMP" occurred frequently in both the abstracts. The literature about Multiple Sclerosis treated inflammation as a concept, as the disease often characterizes

[1]http://www.ncbi.nlm.nih.gov/pubmed last accessed 3 November 2009
[2]http://www.nlm.nih.gov/bsd/pubmed_subsets.htmllast accessed 3 November 2009

inflammations in the central nervous system. The literatures describing the role of MMP9 in Multiple Sclerosis also points out that expression tissue inhibitors of metalloproteinases (TIMP) can control the activity of MMP9 and the relative under expression of TIMP proteins in Multiple Sclerosis conditions. The literature about docosahexaenoic acid also mentions the beneficial effects of the fatty acid on levels of TIMP. So by analyzing the two separate classes of facts mentioned in separate literature sets, it was concluded that fish oil has healthful effects on Multiple Sclerosis (Laine, 2008). Thus by analyzing common concepts and information conveyed these concepts in two different but overlapping fields, it was able to reach a possible conclusion for a question, demonstrating the use of mining free text to find information (see Figure 1.2).

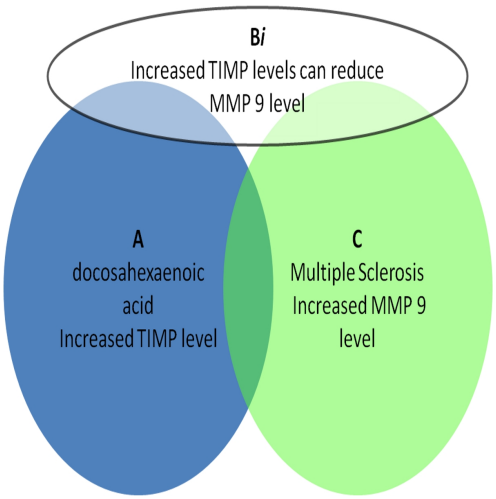


Figure 1.2: ABC model of complementarity adopted from Laine (2008)

But so far, animal genomics has been entirely concentrated on data from databases and data derived through experiments. The purpose of this thesis is the introduction of a text mining tool to the animal science domain, adapting the existing system SCAIView to suit the needs of researchers in the animal science field.

The first chapters of the thesis are focused on animal genomics, particularly cattle and pig genomics, the research directions in these fields, and the use of bioinformatics tools in farm animal genomics field. The follow on chapters are concentrated on problem definition, state-of-the art, text mining, the current status of text mining followed by brief descriptions on ProMiner and SCAIView. The end chapters are dedicated to materials and methods used, and the results and discussions.

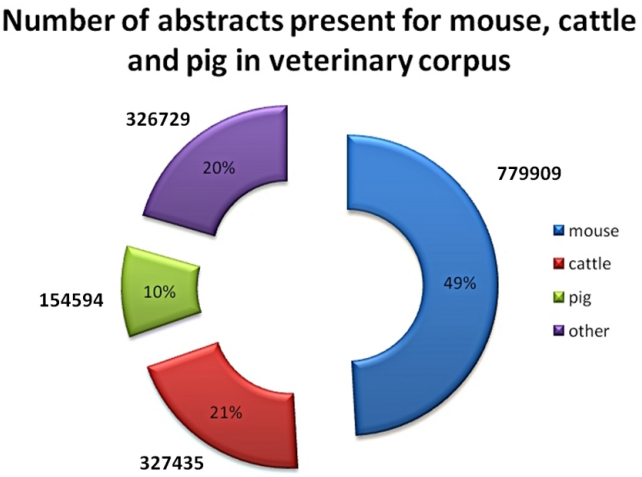


Figure 1.3: Comparison of number of PubMed abstracts for cattle,mouse and pig present in Veterinary corpus: data as of 3 November 2009

2 Livestock genomics: an overview

The origins of modern livestock genomics could be traced back to a series of conferences in the early 1990s, where schemes were developed, and collaborations were established, which aided in maximizing the resources that were available to the livestock genomics in that period. The advances in the human genome initiative paved the way for animal genetics. While animal genomics was still in its infancy, animal geneticists initiated genome projects for some of the most widely used species, which is a reason for the current collection of genomic resources (Womack, 2005). The advances in the genomic studies of farm animals can be categorized into four major sections (1) construction of markers and genes, (2) identifying genes responsible for commercially important traits, (3) use of genome maps to scan across genomes, to identify quantitative trait loci of commercially important traits (4) and all of these finally leading to the production of livestock animals with desirable agronomic (agricultural and economic) qualities (Bulfield, 2000) (see Figure 2.1).

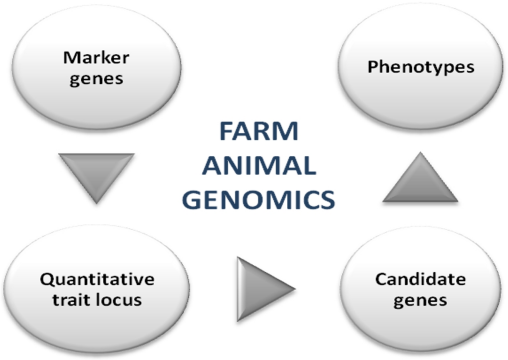


Figure 2.1: Reasearch directions in farm animal genomics

A comparison of human genomics with animal genomics reveals the differences in the research directions of the two fields. In human genomics importance is given to diseases, and the research revolves around genetic diseases, genetic disorders, their prevention, treatment and pharmacogenomics, whereas in case of animal genomics research, commercially important traits are in lime light and attention is given to increasing productivity, better growth yield and rate, and disease resistance.

The advanced knowledge about genomics of behavior, disease susceptibility, morphology and phenotypic diversity of livestock animals combined with advanced and cost effective genotyping technologies have resulted in genomic selection. Advancements in genomic selection could finally lead to marker assisted selection of commercially important animals.

2.1 Cattle and Cattle Genomics: An Overview

Cattle, domesticated even-toed ungulates[1] are raised for meat, dairy products, hides and labour. They were domesticated in the early Neolithic era. Cattle is a ruminant, a mammal of the order Artiodactyla with three fore stomachs and one true stomach.

Table 2.1: Cattle general information

Kingdom:	Animalia
Phylum:	Chordata
Class:	Mammalia
Order:	Artiodactyla
Family:	Bovidae
Genus:	Bos
Number of chromosomes:	30 (28+2)
Taxonomy ID:	9913

Figure 2.2: German Holstein cattle

Figure 2.3: Brown Swiss cattle

There are more than 800 recognized cattle breeds worldwide. Most of the breeds generally fall into two subspecies *Bos indicus* or *Bos taurus*. *Bos indicus* or *Bos taurus indicus*, (also called zebu) are adapted for hot climates. *Bos taurus* or *Bos taurus taurus*, adapted to cool climates, typical cattle found in Europe and are referred to as "taurine" cattle. Hybrids of Taurus and Indicus can be found and are mostly adapted for warmer climates[2]. The cattle breeds present in the Institute of Animal Science Frankenforst research station are German Holsteins (see Figure 2.2) and Brown

[1] Ungulates are animals that use tip of their toes (usually enlarged toe nails called hoof) to sustain body weight

[2] http://www.absoluteastronomy.com/topics/List_of_breeds_of_cattle last accessed 3 November 2009

Swiss[3] (see Figure 2.3). The cattle genome (*Bos taurus*) is fully sequenced as of now and according to Ensembl[4] database the cattle genome has 24,580 genes and has 80% similarity to that of human genome. The bovine genome sequencing project has also shown that human chromosome assembly is more similar to cattle than that of mice or rat. Cattle is the first livestock animal to have its genome completely sequenced and analysed.

Cattle genomics originated from somatic cell genomics, and the first cattle genome maps were synteny groups. Somatic cell genetics and in situ hybridization were combined to assign the synteny groups[5] into chromosomes. An international cattle linkage map was created in the early 1990s and from then on cattle genomics has undergone significant development (Bishop et al., 1994). The development and use of radiation hybrid[6] (RH) maps in high resolution comparative mapping was one of the significant advancements in the cattle genomics field. All these developments led to the final complete assembly of bovine genome. At first, cattle cloning was guided by commercial interest to produce genetically superior animals with commercially desired phenotypic traits. Although the primary objective remains the same, researchers are now using cattle cloning to answer questions in diverse fields such as reproduction biology, developmental and cell biology (Fulka,Jr and Fulka, 2007). Affymetrix has developed microarrays for bovine, called as GeneChip® Bovine Genome Array[7]. The chip was developed through developed through GeneChip® Consortia Program[8] and was built based on Bovine Unigene build 57 data. Bovine Genome Array contains about 24,072 probe sets and they represent approximately 23,000 bovine transcripts and 19,000 Unigene clusters[9]. Illumina has introduced SNP gene chips for cattle in association with United States Department of Agriculture Agriculture Research Service (USDA ARS), University of Missouri and University of Alberta and is called as BovineSNP50 BeadChip. The chip has more that 54,000 SNP probes. The probes are derived from various sources like novel SNPs from Illumina‘s Genome Analyzer, Bovine HapMap data set, and various whole-genome shot gun reads and Holstein (cattle breed) BAC sequence data[10].

[3]http://uf.ilb.uni-bonn.de/versuchsgueter/Frankenforst/de/Betrieb/Tierhaltung/index.html last accessed 8 November 2009

[4] http://www.ensembl.org/index.html last accessed 10 September 2009

[5]Synteny is defined as the co-localization two or more genes are present close together in a chromosome independent of the linkage between them.

[6]Chromosomes are separated from one another and high dose X rays are used to break into several fragments. The order of markers on a chromosome can be determined by estimating frequency of breakage. RH mapping is used to create whole genome radiation hybrid map

[7]http://www.affymetrix.com/products_services/arrays/specific/bovine.affx#1_1last accessed 3 November 2009

[8]http://www.affymetrix.com/partners_programs/programs/consortia.affx last accessed 3 November 2009

[9]http://www.affymetrix.com/support/technical/datasheets/bovine_datasheet.pdf last accessed 3 November 2009

[10]http://www.illumina.com/Documents/products/datasheets/datasheet_bovine_snp5O.pdflast accessed 3 November 2009

Table 2.2: Cattle: genome information
Source: Ensembl database, data as of 02 May 2009

Number of chromosomes:	30 (28+2)
Known protein-coding genes:	20,118
Projected protein-coding genes:	521
Novel protein-coding genes:	686
RNA genes:	2,846
Gene exons:	224,748
Gene transcripts:	30,509
SNPs:	2,057,872

Currently, ruminant biology is making use of the advancements in functional genomics. In addition, developments in comparative functional genomics of cattle genome can also provide significant data on interspecies differences in cell, tissue and organismal biology (Lewin, 2003). Cattle genomics is also focused on producing offsprings with desired phenotypic qualities through various manipulations of embryo, and most of these embryo manipulations are done on embryo at the preimplantation stage. Various keys word search experiments done in PubMed also shows that the number of documents relating to cattle preimplantation in PubMed is also increasing (see Figure 2.5)

Preimplantation stage (see Figure 2.4) is defined as the time period between fertilization and embryo implantation in the uterus, and in cattle preimplantation period is approximately 10 days. This stage is important, in terms of growth and development of the cattle embryo and commercial interests. The beginning of genome transcription of the embryo (embryonic genome activation) and other major morphological and non morphological changes happening in the embryo are crucial to its development whereas commercially, cattle embryos are transferred to the recipient cows when embryos are 7-8 days old at the blastocyst stage. So, any aberrations to the embryo at this stage will adversely affect the proper growth and development of the embryo and commercial interests (Wilmut et al., 1998). In vitro production (IVP) techniques and nuclear transfer (NT) technologies are commonly used for embryo manipulations.

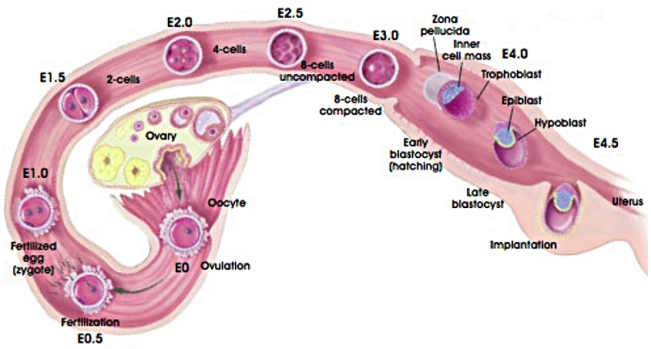


Figure 2.4: Preimplantation embryo development
adopted from http://stemcells.nih.gov/StaticResources/info/scireport/images/figurea3.jpg last accessed 02 October 2009

However, it was noted that there are differences in embryos produced in vivo, and those produced using various embryo manipulation and production techniques. The morphology, gene expression and metabolism of in vitro produced embryos were often different from normal in vivo embryo. In vitro produced embryos also showed problems like increased rated of abortion, large calf size (large offspring syndrome in cattle), aberrant muscle gene expression, increased neonatal mortality and a sex ratio that is skewed towards males (Hansen and Block, 2004). Lower developmental competence, cell number at embryo stages, differences in gene expression, mitochondrial genetics and embryonic, fetal and neonatal problems are accounted for defects and problems related to NT-derived embryos. Researchers in cattle genomics are investigating the genetic reasons behind the problems associated with IVP and NT derived embryos, especially at the preimplantation stage. Hence preimplantation genetics has evolved into a major field in cattle genomics. A list of various genes expressed during cattle preimplantation period in embryos derived from different sources are given in Figure 2.6. A survey of existing cattle genome databases was done for this thesis and is presented in Section 6.1.1.

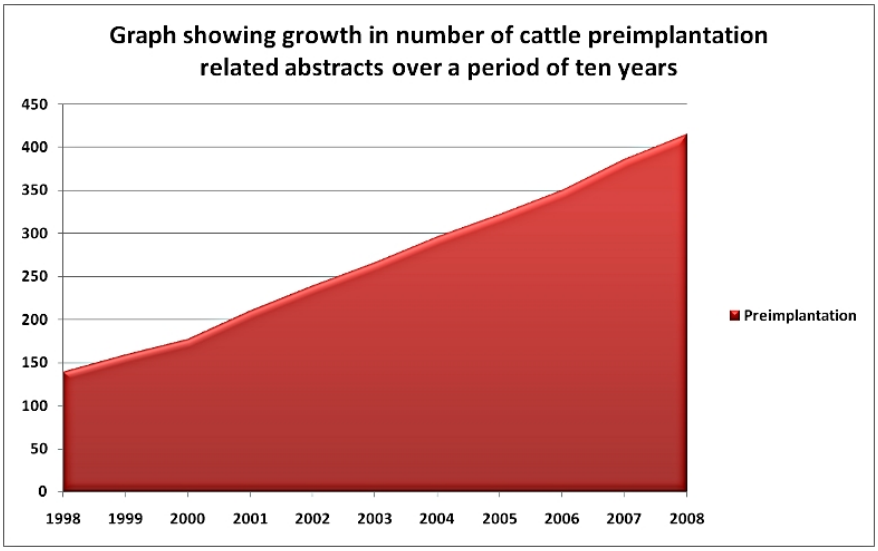


Figure 2.5: Graph showing growth in number of cattle preimplantation related abstracts in PubMed over a period of ten years

Gene categories	Developmental stage	Expression patterns[a] in IVP and sNT-derived embryos compared to their in vivo counterparts		References
		IVP	sNT	
Trophoblastic function				
IF-tau	Blastocyst	Ø/↑	↑	Wrenzycki et al. (2001a, 2001b)
	Blastocyst	↓	n.a.	Bertolini et al. (2002)
Mash-2	Blastocyst	↓	Ø	Wrenzycki et al. (2001b)
X-chromosome inactivation				
Xist	Morula	Ø	↑	Wrenzycki et al. (2002)
	Blastocyst	↑	↑	Wrenzycki et al. (2002); Wrenzycki and Niemann (2003)
DNA methylation				
Dnmt1	Blastocyst	↑	↑	Wrenzycki et al. (2001b) Wrenzycki and Niemann (2003)
Dnmt3a	Blastocyst	Ø	↑	Wrenzycki and Niemann (2003)
Apoptosis				
Bax	Blastocyst	Ø	n.a.	Rizos et al. (2002)
Xiap	Expanded blastocyst	↑/Ø	n.a.	Knijn et al., (unpublished)
Compaction/cavitation				
Cx43	Blastocyst	–	n.a.	Wrenzycki et al. (1996)
	Hatched blastocyst	–	n.a.	Wrenzycki et al. (1996)
	Blastocyst	↓/Ø	n.a.	Rizos et al. (2002)
Cx31	Blastocyst	↑/Ø	n.a.	Rizos et al. (2002)
E-cad	Morula	↓	n.a.	Wrenzycki et al. (2001a)
Dc II	Morula	↓		Wrenzycki et al. (2001a)
	Blastocyst	↑	n.a.	Knijn et al. (2002)
Dc III	Morula	↑/Ø	n.a.	Wrenzycki et al. (2001a)
Plako	Morula	↓	n.a.	Wrenzycki et al. (2001a)
	Blastocyst	↓	n.a.	Knijn et al. (2002); Wrenzycki et al. (2001a)
ZO-1 (pan)	Morula	↓	n.a.	Miller et al. (2003)
	Blastocyst	↓	n.a.	Miller et al. (2003)
Growth factor/cytokine signalling				
LIF	Blastocyst	+	n.a.	Eckert and Niemann (1998)
	Hatched blastocyst	+	n.a.	Eckert and Niemann (1998)
	Blastocyst	↑	n.a.	Rizos et al. (2002)
LR-β	Morula	+	n.a.	Eckert and Niemann (1998)
	Blastocyst	+	n.a.	Eckert and Niemann (1998)
	Hatched blastocyst	+	n.a.	Eckert and Niemann (1998)
	Blastocyst	↑	n.a.	Rizos et al. (2002)

Figure 2.6: List of Preimplantation genes in cattle adopted from Wrenzycki et al. (2004) Legend: ↑up regulated, ↓down regulated, + expression, - no expression, n.a. not analysed, ↑/Ø, ↓/Ø varies with different expression systems

2.2 Pig and Pig Genomics: An Overview

Pig, domesticated even toed ungulate is raised solely for its meat.

Table 2.3: Pig general information

Kingdom:	Animalia
Phylum:	Chordata
Class:	Mammalia
Order:	Artiodactyla
Family:	Suidae
Genus:	Sus
Number of chromosomes:	20 (18+2)
Taxonomy ID:	9823

Figure 2.7: German Landrace pig

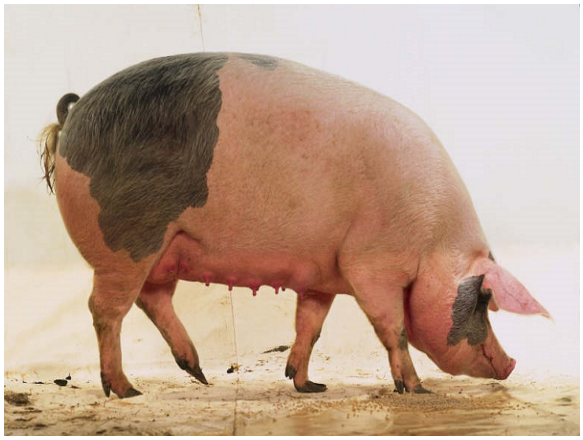


Figure 2.8: Pietrain pig

There are more than 100 pig breeds worldwide and German pig breeds include Angeln Saddleback[11] (Angler Sattelschwein) and Bentheim Black Pied[12] (Buntes Bentheimer Schwein), both of which are considered as rare breeds. Some of the common pig races that are used in the institute of animal science research station are: DL (German Landrace (see Figure 2.7)), DE, Pi (Pietrain (see Figure 2.8)), Du (Duroc), PixDu[13]. Linkage maps for the pig genome were created by European PiGMap initiative and USDA-MARC swine genome project in the midst 1990s. Somatic cell genetic and RH mapping speeded up swine comparative mappings (Womack, 2005).

[11] http://www.ansi.okstate.edu/breeds/swine/angelnsaddleback/index.htm last accessed 3 November 2009

[12] http://www.ansi.okstate.edu/breeds/swine/bentheimblackpied/index.htm last accessed 3 November 2009

[13] http://uf.ilb.uni-bonn.de/versuchsgueter/Frankenforst/de/Betrieb/Tierhaltung/index.html last accessed 8 November 2009

Although many sequencing projects have been undertaken for pig (*Sus scrofa*), the amount of information on pig genes and proteins in the public databases is sparse. A $10 million project is currently undertaken by an international consortium for pig sequencing, with Wellcome Trust Sanger Institute heading the genome sequencing[14]. The pig genome is similar to human genome and has 18 autosomes and 2 sex chromosomes. Similar to bovine gene chips, Affymetrix has developed GeneChip® Porcine Genome Array[15] and is also derived as a part of GeneChip® Consortia Program. The array contains 23,937 probsets for 23,256 transcripts from 20,201 genes. The data for the probesets are compiled from UniGene build 28, Genebank mRNAs till August 2004 and GenBank pig mitochondrial and rRNA sequences. Illumina and International Porcine SNP Chip Consortium have jointly developed a Porcine 60K BeadChip with a total of 62,163 snps on the chip.

Since pig is used as one of the major source of meat, research in pig genomics is also concerned with the genetic factors that affect the quality of porcine meat. In case of meat quality the targets are colour, intramuscular fat, tenderness, pH and water holding capacity. It was found that there are three major commercially important gene effects for porcine meat quality, such as:-

- **The sex chromosome effect**: the quality difference in meat between barrows (castrated male hog) and gilts (female hog). The difference is that barrow meat has a marbling effect[16] compared to gilt meat and barrows have more back fat and less meat yield percentage when compared to gilts.
- **Stress gene effect**: which was first described as porcine stress syndrome. It was found that when some hogs were stressed physically, were more susceptible to death and produced a pale soft and exudative meat, a condition referred to as Porcine Stress Syndrome (PSS). Meat pH, colour, drip loss, intra muscular fat and tenderness are deleteriously affected by the HAL gene.
- **Napole effect**: the condition caused by the Napole (RN-) gene, the result of which is a low muscle pH, cooking loss and water holding capacity, yet, Napole gene has a positive effect on meat tenderness. The state caused by Napole gene is mentioned as the Hampshire effect.

In the quest for improved meat quality, researchers in pig genomics are working on two separate directions. In the first direction, researchers are working on identification of major genes and polymorphisms in them, which are responsible for the basis meat qualities. In the second direction, scientists use DNA marker information to identify the stretch of DNA that is closely linked to the gene responsible for a target trait (QTL). This method is addressed as Quantitative Trait locus Mapping (QTL mapping) (de Vries et al., 1998). A survey of existing pig genome databases was done for this thesis and is presented in Section 6.1.2.

[14]http://www.sanger.ac.uk/Info/Press/2006/060116.shtml last accessed 15 July 2009

[15]http://www.affymetrix.com/products_services/arrays/specific/porcine.affx last accessed 4 November 2009

[16]the presence of intramuscular fat in red meat giving the appearance of a marble like pattern in meat

2.3 MicroRNAs: An Overview

"MicroRNAs (miRNAs) are endogenous ∼22 nt RNAs that can play important regulatory roles in animals and miRNAs plants by targeting mRNAs for cleavage or translational repression" (Bartel, 2004). The microRNAs were initially discovered in *C. elegans*, (Grad et al., 2003) in which two 22 nucleotide long noncoding RNAs were found to interfere in growth and regulation. The *let-7* RNA was found to promote the transition from late larval to adult cell fates and second on *lin-4* RNA was found to act in the development to promote progression from the first larval stage to second. Experimental and computational methods have helped in the identification of miRNAs in most of the organism including plants. In different organisms they function differently, in flies miRNA functions include control of cell proliferation, cell death and fat metabolism, where as miRNAs have a role in neuronal patterning in nematodes and in mammals the function include modulation of hematopoietic lineage differentiation. In plants, they control the leaf and flower development.

MicroRNA biogenesis (see Figure 2.9) begins with the transcription of the miRNA gene by RNA polymerase II, which generates a long primary miRNA that contains mature miRNA as RNA hair pin (see Figure 2.10). A 70 bp precursor miRNA is formed as the primary miRNA is cleaved by Drosha, an endonuclease. Exportin-5 protein transfers the precursor miRNA from the nucleus to the cytoplasm where it is further processed by an endonuclease enzyme called as Dicer. In this step, the ∼21 nucleotide long hairpin loop of miRNA is cut. The resulting miRNA:miRNA*(miRNA* is the complementary microRNA strand) duplex is identified by RISC(RNA Induced Silencing Complex) and the microRNA complementary strand is generally degraded. The resulting complex functions as mechanism of post transcriptional control. The miRNA attached to RISC targets specific binding sites of mRNA is the 3' UTR (untranslated region) (Kim, 2005). RISC is a ribonucleoprotein complex needed for the miRNA mediated gene silencing. The component proteins of the RISC include agronuate proteins (AGO1 to AGO4) (Rand et al., 2004), the Dicer protein and human immunodeficiency virus-1 transactivating response element RNA binding protein.

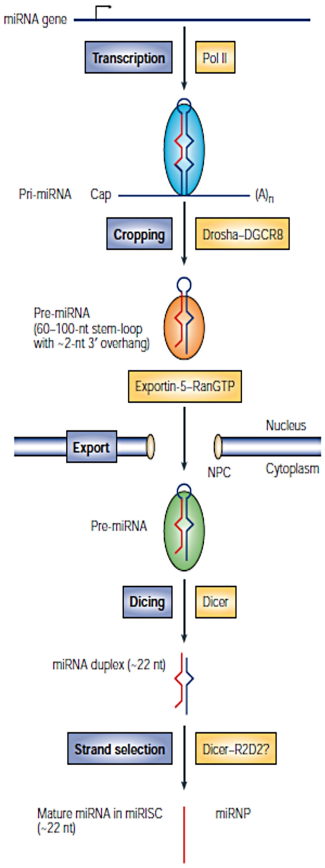


Figure 2.9: MicroRNA biogenesis adopted from Kim (2005)

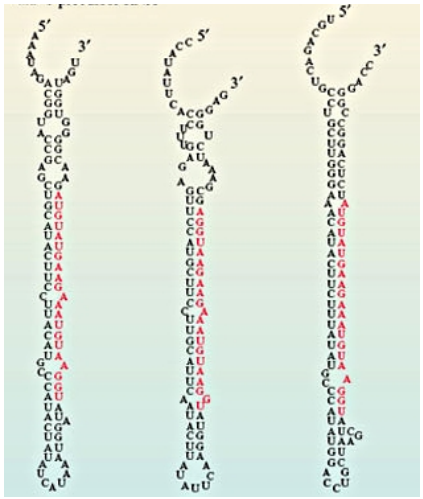


Figure 2.10: MicroRNA hairpins adopted from http://openlearn.open.ac.uk/file.php/2645/S377_1_019i.jpg last accessed November 2 2009

Three major theories have been postulated on the post transcriptional gene regulation mechanism by miRNA (McDaneld, 2009). The first theory proposes that the target mRNA is endonucleotycally cleaved at the binding site by miRNA. This theory proposes that miRNA targeting of mRNA is based on number and type of base pairs that match at the 3' UTR binding site of miRNA (Engels and Hutvagner, 2006). A second theory proposes blocking of initiation by miRNA:RISC complex. This theory proposes that although the miRNA:RISC complex binds to the 3' UTR region of miRNA the post transcriptional gene regulation is achieved through a cascade of events that finally block the initiation proteins from binding to the 5' cap of mRNA (Chendrimada et al., 2007). A third theory proposes that miRNA:RISC complex transports mRNA to P-bodies (Chan and Slack, 2006). P- bodies are regions in cytoplasm which are rich in enzymes and factors for mRNA turnover and repression of translation but lacks ribosomal proteins for translation. P bodies are believed to be site of miRNA action (McDaneld, 2009). A noteworthy characteristic of animal microRNA is that the genes coding for microRNAs in animals are grouped and clustered together in genome and these clustered miRNAs are produced from a single mRNA molecule resulting from the translation of several genes. It was also found that if the produced miRNA molecules have sequence similarity, the association might contribute an additive effect to gene regulation. For example, the miRNA genes *miR-125* and *let-7* are clustered together in fly genome and these are regulated together (Ambros, 2004). Figure 2.11 gives a graphic representation of various theories proposed for microRNA gene regulation.

The advancements in miRNA research has led to the research of miRNA for traits of economical importance in farm animals. The first subject of interest was the factors that affect the development and growth of economically important tissues like skeletal muscle and adipose tissue. There were instances pointed out where several miRNAs were found to have a direct influence on myogenesis and associated pathways and certain muscle specific miRNAs that has a direct influence on economic traits in livestock were also reported (McDaneld, 2009). A mutation in the myostatin gene of the heavily muscled Belgian Texel sheep creates a target site for the microRNAs *miR-1* and *miR-206* in the 3' UTR region of the transcript. This mutation results in a decreased translation of the myostatin protein and a resulting increase in muscle mass. In addition to miRNAs affecting muscles, research was also done in miRNAs having an effect on adipose tissue. In human pre adipocytes, *miR-143* was found identified; the amount of the same increased during adipocyte differentiation and it was noted that adipocyte differentiation rate was decreased upon the inhibition of the miRNA (Esau et al., 2004). Transcriptome profiling of miRNAs were done to determine the role of miRNAs in livestock species. It was found that a large proportion of miRNAs were expressed in all cell types and a certain number of miRNA were expressed in certain cell types and during certain developmental stages of the embryo. The emerging areas of microRNAs research in livestock genomics include reproduction, immunology and feed efficiency. Recent developments in this field include identification of bovine miRNAs expressed in cumulus oocyte complexes during late oogenesis (Miles et al., 2009) and determining the miRNA motifs that are associated with genetic variants that are responsible for different residual feed intake
(Barendse et al., 2007).

Table 2.4: Number of microRNAs identified in cattle, human and mouse: data as of 28 Septemer 2009, source miRBase version 14.0

Organism	Number of microRNAs identified
Cattle	356
Human	706
Mouse	547
Pig	77

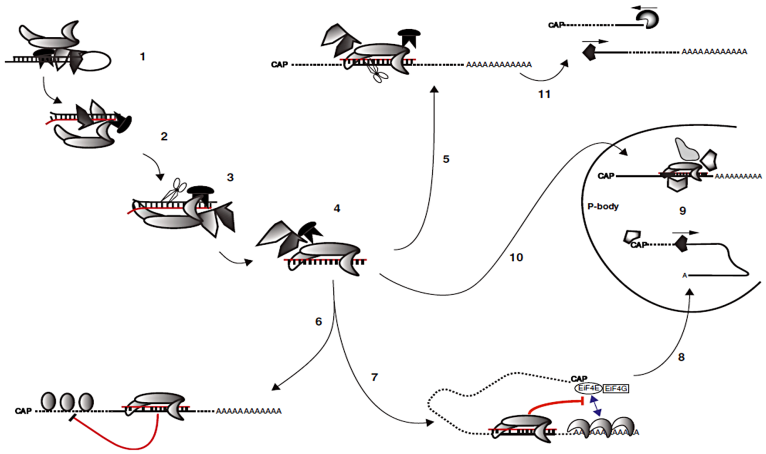


Figure 2.11: MicroRNA RISC assembly and functions adpted from Engels and Hutvagner (2006)
Legend: (1-4) miRNA complex formation with RISC, (5) endonucleolytic cleavage of mRNA, (11) mRNA further degradation, (6) Blocking translation initiation, (7-9) mRNA excluded from translation and transported to P-bodies storage and degradation, (10) RISC assembly

3 Knowledge discovery and text mining in biomedical domain: an overview

3.1 Sources of biomedical literature

With the onset of online publication, literature published anywhere in the world has become accessible to anyone in any part of the world. Especially the Biomedical domain saw a boost in the number of literatures published. Online publishing also helped scientists to understand what their counterparts on other side of the globe is interested in, on a daily basis. Although the overall amount of data accessible has been increased tremendously, the reliability of these data is at question. But, this problem can be alleviated by using reliable online data sources. If analyzed from a biomedical framework, the largest data source in the World Wide Web is PubMed, a service of the United States National Library of Medicine (NLM[1]). PubMed, along with its largest component MEDLINE, covers over 19 million citations and some full text articles in PubMed Central. Other major players in this field include Nature Publishing Group[2], Science[3], Science Direct or Elsevier[4] who provides access to high quality journals based on a licensing fee system. At this point, it is worth mentioning about some of the other publishers like Oxford Journals, which also has high screening standards for the journals selected for publishing. Now, the major sources of biomedical literature, PubMed and MEDLINE will be explained in detail.

3.1.1 MEDLINE (Medical Literature Analysis and Retrieval System Online)

MEDLINE is a leading bibliographic database containing literatures from the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences by United States National Library of Medicine (NLM). Literature Selection Technical Review Committee (LSTRC) gives recommendations for journals to be selected for MEDLINE. Currently MEDLINE holds more than 19 million citations from 5,200 worldwide journals in 37 languages. MEDLINE comes under PubMed, which is a

[1]http://www.nlm.nih.gov last accessed 2 November 2009
[2]http://www.nature.com last accessed 10 November 2009
[3]http://www.sciencemag.org last accessed 10 November 2009
[4]http://www.us.elsevierhealth.com/index.jsp last accessed 10 November 2009

part of Entrez series of databases maintained by NLMs National Center for Biotechnology Information (NCBI). A unique feature of MEDLINE is that the records in MEDLINE are indexed by MeSH (Medical Subject Headings) controlled vocabulary by NLM. Since its introduction in 1971, MEDLINE quickly gained popularity. Since 2005 around 2000-4000 completed references are added each day, Tuesday through Saturday. Figure 3.1 shows the number of PubMed searched made in a span of ten years, showing the increasing popularity of PubMed.

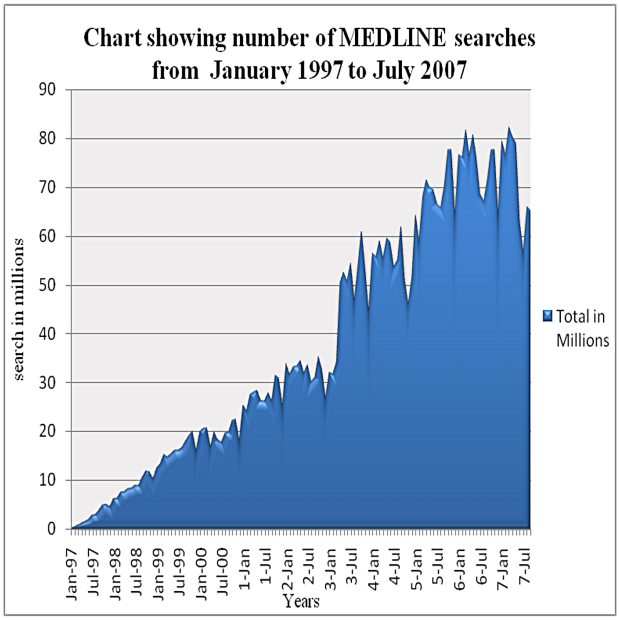


Figure 3.1: Number of MEDLINE searches made from Jan. 97 to July 07

MEDLINE search engines are designed to use a Boolean expression that combines MeSH terms, keywords in abstract and title of the article, author names, published date and so on. MEDLINE search engines allow also querying for similar abstracts depending on a mathematical scoring function which takes into account the word content of the abstract and title of the article[5].

3.1.2 PubMed

PubMed is the search engine that is primarily used to access MEDLINE, which is built to search MEDLINE for abstracts, citations and full text articles in fields relating to biology and medicine. Some of the services provided by PubMed include free access to MEDLINE, links to full text article sites and related information sources, links to PubChem and many other molecular biology databases, links to articles related to a selected citation and so on. In addition to MEDLINE, PubMed also contains the following:

[5]http://www.nlm.nih.gov/pubs/factsheets/medline.html accessed September 13, 2009

- In process citations provides a record of an article before the article is indexed with MeSH or changed as out of scope article.
- Some of the OLDMEDLINE citations which are not updated with the present vocabulary.
- Out of scope citations from certain general science and general chemistry journals.
- Some full text articles that are submitted to PubMedCentral, but not recommended for inclusion in MEDLINE.
- Some early physics journals which were a part of the prototype in the early 1990s.
- Some citations that were done before the journal was subjected for MeSH indexing[6].

A detailed analysis of the difference between information content of full texts in PubMed Central and abstracts in PubMed documents was done (Mueller, 2009). In the study it was found that there are approximately more than 1.89 million full text documents in PubMed Central. As a part of this thesis full text documents in PubMed Central for cattle and pig were searched and it was found that PubMed Central contains 1363 cattle related full texts and 643 pig related full texts.

When a search query is submitted in PubMed, PubMed will try to match the terms in the query to a series of lists through a feature called as Automatic Term Mapping. PubMed will try to match the terms in the search query to subject in the MeSH translation table (an alphabetical list of MeSH terms, subheadings, references, names of substances and synonyms). If the term was not successfully matched to a MeSH term Automatic Term Mapping will try to match the query as a journal in the Journal Translation table or as Author and Investigator names in Full Author translation table, author index, Full Investigator translation table or investigator index. If PubMed is not able to match the phrases, then those are broken apart and the process is repeated. PubMed also matches the terms and phrases to a stopword list (a list of the most commonly occurring words). These terms are not included in the search or indexing, as a search with these terms would return almost all of the articles as a hit.

PubMed also supports search using Boolean operators (AND, OR, NOT), a search in PubMed using Boolean operators requires the use of Boolean operators in capital letters. PubMed supports nesting of Boolean operators. Normally Boolean operators in a search are processed from left to right. When search terms along with Boolean operators are enclosed in parentheses, these are processed as a single unit. For example a search query like "cattle preimplantation AND IVF OR cloning" will retrieve all the documents citing cattle preimplantation period and IVF along with all the documents with the term cloning in it (see Figure 3.2). If the same terms and Boolean expressions are used with parentheses like "cattle preimplantation AND (IVF OR cloning)", the result would be different(see Figure 3.3).

[6]http://www.nlm.nih.gov/pubs/factsheets/pubmed.html accessed September 13, 2009

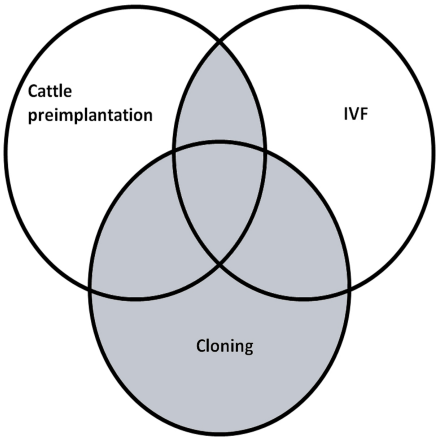


Figure 3.2: PubMed un-nested search

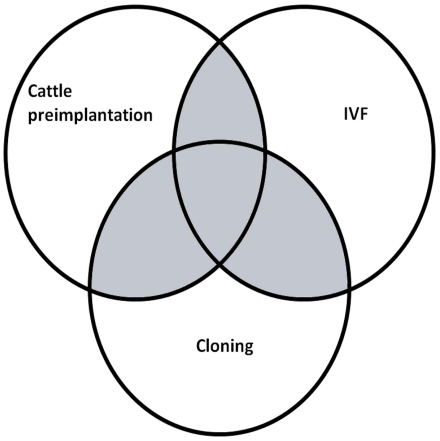


Figure 3.3: PubMed nested search

This versatility of the PubMed search engine is the main reason why most of the text mining approaches in biomedical domain to use PubMed as one of the prime search engines to retrieve text.

3.2 Knowledge discovery

Knowledge discovery is a term that is used in close association with data mining and text mining. Knowledge discovery is defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley et al., 1992).

The proposed knowledge discovery process consisted of nine stages which were summarized as (Fayyad, Piatetsky-Shapiro and Padhraic 1996):

- **Developing and understanding the application domain.**
  The first step, where the needs of the end user are understood and basic knowledge about the domain is gathered.
- **Creating a target data set.**
  This step involves querying the already existing data set to select the target subset and sampling variables (attributes) and data points (examples) that are to be used to perform knowledge discovery task.
- **Data cleaning and preprocessing.**
  The tasks such as removing the outliers, noise reduction and dealing with missing values are done at this stage.
- **Data reduction and projection.**
  This step includes processes like data transformation and data dimension reduction which would finally the projection of data in its homogeneous form.

- **Choosing the data mining task.**
  Here decision is made on the data mining task (like classification, clustering or regression) to be used, so that the objectives defined in the first step are met.

- **Choosing the data mining algorithm.**
  The data mining task selected in the previous step is put to use at this phase. The models used and parameters applied in the method are adjusted to produce an optimum result.

- **Data mining.**
  The patterns are generated at this stage. For example, patterns like classification rules, decision trees, regression models and trends etc.

- **Interpreting mined patterns.**
  The extracted patterns and models are visualized and the visualization of the data is based on the models that are extracted.

- **Consolidating discovered knowledge.**
  The final step, where the discovered knowledge is incorporated into the performance system and is also done at this stage.

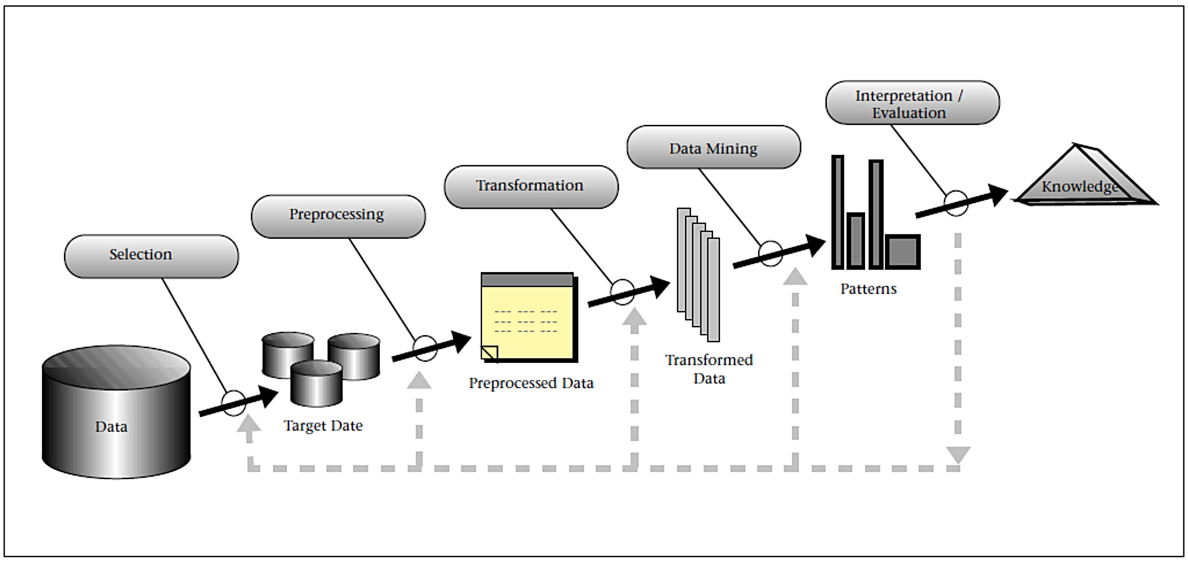


Figure 3.4: Knowledge discovery process
Adopted from http://liris.cnrs.fr/abstract/fayyad1996.png last accessed 02 October 2009

The process is iterative and at each and every step a change of parameter or further refinement of the data would result in an altered knowledge representation (Figure 3.4).

3.3 Text mining

Text mining can be summarized as the process of extracting previously unknown information from different written sources, and linking together the newly derived information to frame new hypothesis or facts. In the biomedical domain, text mining is primarily used to recognize biological entities through named entity recognition. Text mining in biological domain is involved mainly in recognizing biological entities such as genes and proteins in written text, extracting protein-protein interactions through automated processes and mapping proteins to their functions. From the very beginning, applications of Information Extraction systems in the biomedical world have been concentrated on Medline abstracts. Recent developments point out that text mining in biomedical domain also involves identifying drug names, protein drug interactions, host pathogen interactions and units important in the biomedical domain. From a text mining point of view, Knowledge discovery can be redefined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data by the analysis of complexities of textual data.

Biomedical text mining started in the late 1990's with gene protein name identification and protein-protein interaction detection. In a biomedical background, **named entity recognition** (**NER**) can be outlined as identification of biomedical entities like genes, proteins, drugs, enzymes and so on. A prime source for NER in biomedical domain would be dictionaries containing the entities and their synonyms used in the domain. An error at this step can lead to problems in later stages. For instance, in a protein and gene name dictionary, the entries need to be unambiguous and uniquely mapped to a database identifier (like Entrez Gene or UniProt). Given below is a classical case of gene name ambiguity:

In many abstracts, the term "GC" is used to refer to the gene named "Group specific Component". The same term GC is used to cite "Guanylate Cyclase" in a lot of other instances and finally, GC is used while mentioning the occurrences of the nucleotides "Guanine" and "Cytosine" (GC rich region in a gene, for example). Term normalization is the name given to the process by which ambiguous gene and protein names are disambiguated and mapped to a unique database identifier. The outcome of disambiguation process depend on the organism under consideration, the quality of gene annotation and the amount of gene and protein data that is present in public databases (Entrez Gene, UniProt) for the organism (Erhardt, Schneider and Blaschke 2006). Biomedical knowledge discovery/text mining has special importance, given the fact that the number of scientific journals in the biomedical context is increasing steadily and not all of the hidden information in texts makes it to the public domain databases. For example, it would be interesting for the researchers in the animal science domain to capture information about a previously unidentified protein-protein (gene-gene) interaction that is causing structural differences of an artificially transferred embryo, which is hidden within the latest research publications.

Traditional text retrieval or information retrieval method comprises of formulating search queries and submitting them to the search engines. Here the quality of the returned hit depends on the expertise of the user in formulating search queries. Normal text

based search engines sees the document as a ‘bag of words‘ and index words associated with each document. Based on the algorithms in search engines, the user formulated queries are related to the indexed terms and relevant hits are returned. Still the end result of text searching is a collection of documents, where the information is hidden in plain sight from the user. Text based knowledge discovery can be regarded as a method of refining the user query, retrieving and visualizing the relevant information in the documents and using the patterns in the retrieved data to present new information and frame hypothesis. Using text mining methods can improve the search quality, assist in information extraction and leads the way to knowledge discovery from text.

Normal search methods are highly influenced by the search pattern and search behavior of the user, meaning that the results returned depend on how user formulated the search query. There are methods in existence, which would improve the search. One of the methods is the query refinement method. This method works on the concept that the search terms submitted by the user are not the best ones to express the information as these terms do not match the terms expressed in the documents. The purpose of the query refinement method is to assist user in distinguishing those terms which would probably appear in the documents. One classical example is the feedback system, where user has an option to retrieve more documents like the one which has best matching to the query terms. The second approach, natural language searching gives user more freedom in expressing the search query. These methods allow users to submit queries like “What are the genes that are expressed in the 32 cell stage of the cattle embryo“. These methods extract and index the semantic structure of the query terms and also give importance to relationship between terms. The third method is the clustering of documents. The clustering is based on content similarity of the documents explained by prominent terms, which are shared by all the documents in the cluster. The next method is document categorization, where documents are ordered by separating them into different categories. Before document categorization can be done for a particular field, a domain expert should define and name the categories that make up the field. The final method is the document summarization method, where an abstract or summary of a full text document is generated automatically. Different techniques are used for document summarization; some involve extracting the keywords from a document and presenting these keywords as a summary, some others include scanning for significant terms in a document and presenting the sentence containing those terms as summary. Certain sophisticated techniques modify the sentence selected from different parts of a document to produce a uniform text.

Information extraction methods in text mining are aimed at the automated identification of entities, in this case biological entities such as gene/protein names, enzymes, drugs and the relationship between those terms. The enhanced search methods explained in the previous section (refer to Section 3.1.2) also extracts the key entities in the text, but information extraction is more focused towards extracting entities and entity relationships that expresses a fact. Two different kinds of relationship modeling exist in biomedical text mining. The first one is the co-occurrence based relationship mining, where relationship is established between two entities if they occur together. The next one involves the use of natural language processing (NLP) methods. NLP methods involves use of natural

language grammar based rules and statistical methods over text to define relationship between entities. Visualization of extracted terms and relationship is also an important part of information extraction (Robert and Michael, 2002).

Information extraction can enrich the text based understanding by identifying relevant entities and relationship between them and provides a better way of understanding the data in the text. But, researchers in the biomedical domain are interested in acquiring novel facts about biological entities than knowing about the entities and entity relationship in a document. Knowledge discovery and knowledge environment representation gain importance in this context. Knowledge discovery in a biomedical angle concerns with associating and assembling textual descriptions of data along with other data forms (data from biological databases, experimental data etc) to create a descriptive medium that could help in understanding the meaning and significance of data. The biological database, UniProt can be modeled as a knowledge environment representation. It integrates data from multiple sources. The UniProt database itself, data from GO annotation (functional gene annotation), Interpro database for protein family classification accounts for some of the database information in UniProt, where as matched text sentences, PubMed ID and comments on the protein are associated textual information present in the database. The data from experiments are present in the form of some of the protein sequence information, confirmation of existence of protein and method of protein crystal study. Another example for a knowledge environment representation is SCAIView, explained in the material section of materials and methods chapter.

3.4 Named Entity Recognition Systems

Based on the algorithmic approach used, Named Entity Recognition systems can be sorted into different approaches. They are:

- **Dictionary based approach**: Method of matching an entry from a large collection of names (dictionaries) against text. This approach also includes fuzzy matching or inexact matching to include spelling variants, use of Greek numerals and use of hyphen (-) instead of space.
- **Rule base approach**: Separation of different entity classes using a set of custom designed rules. For example, use of term topology information like use of capital letters, symbols and digits to identify gene and protein names.
- **Classification based approach**: The systems making use of this approach reduces the task of NER into classifying entities into different classes. In this case, into different biological classes like genes, proteins, enzymes and so on.
- **Sequence based approaches**: Sequence based approaches consider the order of words and phrases into account, and performs a statistical analysis to figure out the most probable sequence of words and phrases for a given set of words.

3.5 NER Performance evaluation

The performance evaluation of named entity recognition systems are done by calculating precision, recall and F-score (F_1 measure) of the system. Precision of a system is defined as the proportion of relevant entries (for example, genes) in the entries retrieved in total. True positive is defined as an entity identified by both the 'gold standard' considered and the system under evaluation. False positive in a system is defined as the entity found only by the system under evaluation, not by the gold standard, that is an entity that is incorrectly identified by the system under evaluation. False negative is defined as an entity that is found only by the gold standard, not by the evaluated system, that is an entity missed by system under evaluation. F-score is defined as the weighted average of precision and recall or in other terms the harmonic mean of precision and recall. Maximum value for F-score is 1 and the least is 0.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ positive} \tag{3.1}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ negative} \tag{3.2}$$

$$F_1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.3}$$

Since F_1 measure is defined as the balanced harmonic mean between Precision and Recall, it can also be written as:

$$F_\beta\ score = \frac{(\beta^2 + 1)\ Precision \times Recall}{\beta^2\ Precision + Recall} \tag{3.4}$$

for a non negative real β, where β is used as a parameter to control the relative weights that are given to precision or recall. Here, F_β is defined as "F_β measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision" (Rijsbergen, 1979).

All these calculations are done by comparing the results of the system to a gold standard. In case of system for gene and protein identification, the gold standard is a corpus with its genes and proteins annotated by an annotator.

BioCreAtIvE: Benchmarking biomedical text mining

For the evaluation of text mining systems and information extraction systems in Biology, a community wide effort called the BioCreAtIvE[7] (Critical Assessment of Information Extraction systems in Biology) has been initiated. Text mining approaches in biology were addressing different problems and the performance evaluation of each method was done using private data sets. BioCreAtIvE was introduced as a performance measure to determine the quality of existing biomedical text mining tools in biomedical domain, to test their real world applications and to determine the performance.

[7]http://biocreative.sourceforge.net last accessed 2 November 2009

There were two BioCreAtIvE challenges so far, first one in 2004 and a second one in 2006. In BioCreAtIvE 2003/4 the challenges were the identification of gene mentions in text and linking protein database entries to abstracts and extraction of human gene product with GO terms. The focus of BioCreAtIvE I (2003/4) was on model organisms fly mouse and yeast. In 2006/7 the tasks were gene mention tagging, gene normalization and extraction of protein-protein interactions from text. The second BioCreAtIvE (BioCreAtIvE II) challenges were focused on human genes.

3.6 State-of-the-art dictionary based approaches

3.6.1 ProMiner

ProMiner is a dictionary and rule based Named Entity recognition system developed at Fraunhofer Institute for Algorithms and Scientific Computing SCAI. ProMiner has been successfully tested and is being used along with the associted knowledge environment platform SCAIView a as a Named Entity Recognition tool and knowledge recognition platform in human, mouse and Arabidopsis genomics and associated fields. ProMiner and SCAIView are explained in detail in Section 7.1.11.

3.6.2 AliBaba

AliBaba is a dictionary based NER system for recognizing biomedical objects developed at the Humboldt University, Berlin. AliBaba was developed as an interactive graphical tool that graphically summates the search results. AliBaba is designed to extract associations between cells, drugs, proteins species and tissues. The dictionary used in AliBaba is gathered from different databases providing information on proteins and genes. Along with the dictionaries, AliBaba also uses regular expressions to account for spelling variants. The algorithm in AliBaba uses two different approaches are used in parallel: pattern matching and co-occurrence filtering. AliBaba uses the former to extract protein-protein interactions and protein cellular locations. The patterns in the pattern matching algorithm are extracted from annotated task specific corpora and are made up of regular expressions using tokens, part of speech tags and entity classes. Based on the match quality between the sentence and the pattern, pattern matching algorithm provides a confidence score for each relation. The core of AliBaba is an information extraction pipeline, named as IE Pipeline. IE Pipeline holds a sentence splitter, a tokenizer, a part-of-speech tagger (POS), a stemmer, and a dictionary based named entity recognizer as several modules(Palaga et al., 2009).

AliBaba is designed as a client server application for the extraction and graphical visualization of biological entities (see Figure 3.5). AliBaba client is a Java Web Start application which accepts queries exactly in the way PubMed does. The web interface of AliBaba also allows limiting the number of citations to a user specified number. When a search query is given in the web interface of AliBaba, the query is forwarded to PubMed, the result of which is PubMed identifiers (PMIDs) of all the matching abstracts.

The algorithm in AliBaba then searches for abstracts of the respective PMIDs in the AliBaba server. The PMIDs which are not found in the server are sent back to the PubMed to retrieve the abstracts, which are then sent back to the AliBaba server. The IE Pipeline in AliBaba processes the abstracts, which are then sent back to the client. The client enriches the abstracts with external database information and displays those (Plake et al., 2006).

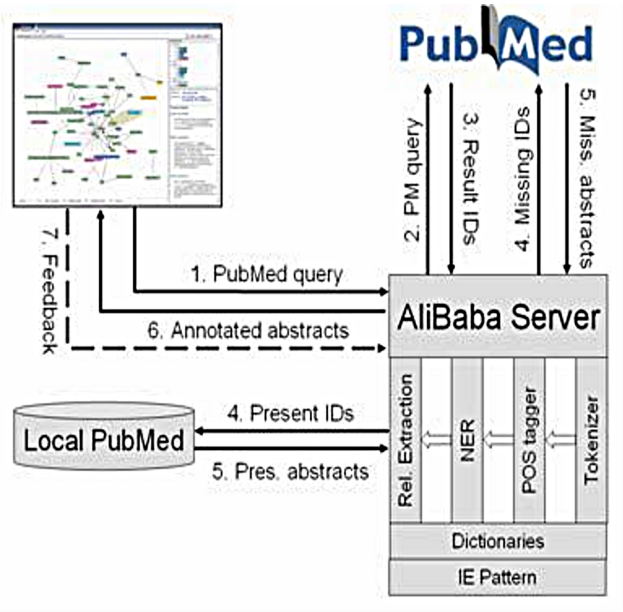


Figure 3.5: Alibaba workflow adopted from Plake et al. (2006)

The client screen of AliBaba consists of three regions (see Figure 3.6), the query field which accepts search queries to PubMed being the first. The large window right below the input field shows the graph resulting from parsing the abstracts obtained as hits of the query. The nodes of the graph, representing the biological entities are colour coded. The edges of the graph represent the association between two different entities. The right hand pane next to the graph window consists of two tabs, the first one called "Objects" and the next one called "Texts". The identified biological entities are grouped into different classes of biological entities (proteins, enzymes). Clicking on an identified entity in the pane gives the relevant PubMed abstract. The second tab "Texts" consist of a list of titles of the entire retrieved PubMed abstract and clicking on a title gives the full abstract with the biological entities colour coded. AliBaba also provides link out to UniProt.

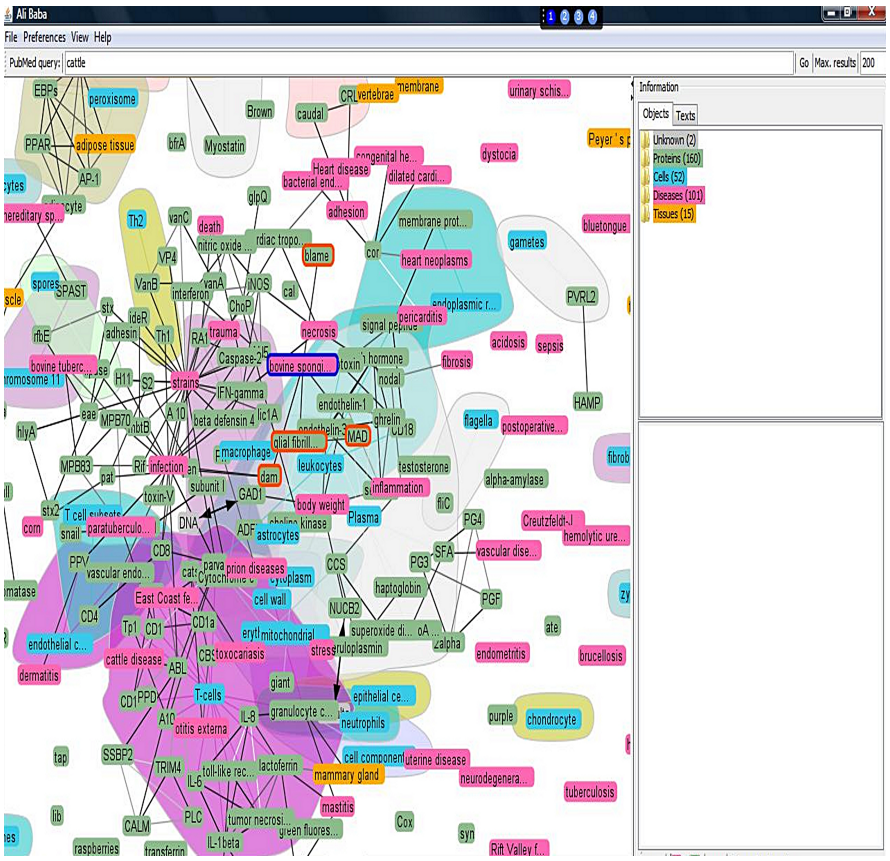


Figure 3.6: Alibaba search result view

The extensive coverage of AliBaba makes it more or less impracticable for organism specific applications, a search for cattle proteins retrieved proteins from cat, pig, rat and orangutan. There are problems with the mapping of external database information as well. For example, AliBaba identified HAMP as a cattle protein, but the linking to the UniProt provided information on HAMP protein in zebra fish (UniProt accession Q7T273), human (UniProt accession P81172), Pig (UniProt accession Q8MJ80) and not on cattle HAMP protein. AliBaba do not provide link outs to other important databases such as Entrez Gene and GO annotations of the identified gene or protein is also not available.

3.6.3 EBIMed

EBIMed is a web application by European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus. EBIMed extracts abstracts in the same way as PubMed does and provides a service in which co-occurrence based analysis of the MEDLINE abstracts is associated with document retrieval. The biological entities under consideration are gene/protein names, functional annotations of genes and proteins (GO terms), drug names and species names. Based on co-occurrence of these biological entities EBIMed

maps protein-protein interactions, functional annotations of genes and proteins, drug-protein interaction and does the categorization of proteins as proteins of model organisms. The biological entities for the co-occurrence based system in EBIMed are protein names and synonyms from UniProtKB/Swiss-Prot, GO terms, drug names from MedlinePlus and species names from NCBI taxonomy which are provided to the system as a list of synonyms. The MEDLINE abstracts provided by NLM are indexed with Lucene. The current indexing covers title abstract text, author list and MeSH terms. Gene and protein name normalization is done as a part of tokenization in indexing. Token normalization converts the gene/protein names into lower case characters and splits characters from digits. In the normalization step, irregular verbs are reduced to the base form and plural forms are converted to singular.

The process pipeline of EBIMed contains certain modules, which performs the tasks in a step by step manner. Xml is the input and output file format for these modules. The identification of genes and proteins is based on an xml file containing all the UniProt/Swiss-Prot protein names and synonyms. The identification of genes and proteins follow certain set of rules, and optional characters (‘-‘,‘_‘,‘/‘) are allowed in protein names instead of blank spaces (‘ ‘) (Rebholz-Schuhmann et al., 2006). This allows the system to identify the different variants of a protein name. For example, the said rule allows the identification of ‘IGF 1‘ and ‘IGF-1‘ as the same protein. The found acronyms are marked if the expanded form is found in parallel, or else the the acronyms are omitted from marking. The protein names identified are tagged and linked to the corresponding entry in the UniProt database. An approach similar to the identification of proteins is followed for GO terms. For the co-occurrence mapping, all the sentences in which contain a pair of terms (identified biological entities) are gathered, sorted and clustered. The pairs are ranked according to the highest number of evidence sentences present and are listed in a descending order according to the number of evidence sentences present.

The top most section of the result page for EBIMed provides a link to all the retrieved abstracts and table of all the biological entity classes with the statistics for number of hits and hit pairs found. The bottom section is divided into columns in which the first column contains the identified gene/protein names. The names provide a hyperlink to a page which shows the sentence in every the abstracts where in which the particular protein was identified. This page also provides the PMID of the abstract. The second column gives the name of all the proteins/genes which co-occurred with the protein. The next three columns give information about the functional gene annotation (GO terms) of the protein, followed by the column which gives information on drugs. The last column shows the model organisms in which the particular protein was found (see Figure 3.7).

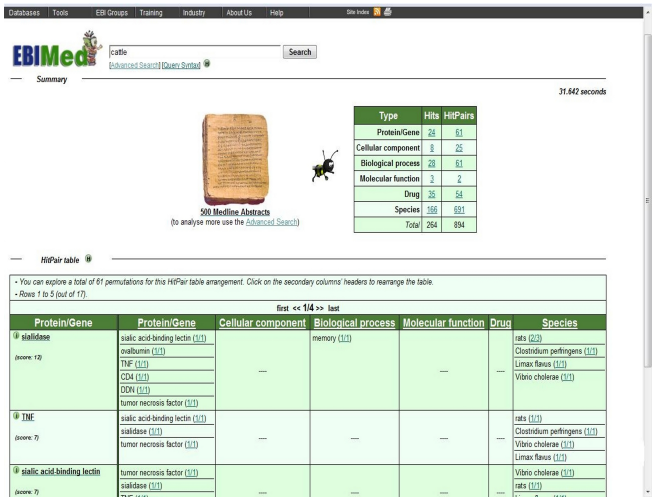


Figure 3.7: EBIMed search result view

3.6.4 Information Hyperlinked over Proteins (iHOP)

iHOP is a NER system developed at the National Center of Biotechnology. iHOP is built on the principle that the entire PubMed can be transformed into one navigable source when genes and proteins are used as hyperlinks between them and knowledge discovery through overlaying the literature network over experimental data (Hoffmann and Valencia, 2005). The present version of iHOP is designed to work on data from a handful of mammalian species, other medically important entities and model organisms and plants such as Arabidopsis and rice. iHop is implemented as a dictionary based approach. The entire architecture of iHOP (see Figure 3.8) is divided into two parts, the iHOP factory and the iHOP web application.

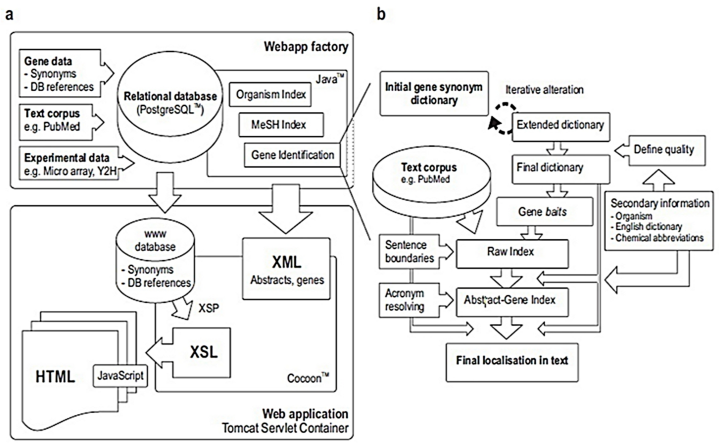


Figure 3.8: iHOP architecture adopted from Hoffmann and Valencia (2005)

The first section, the iHOP factory is responsible for source data management and manipulation, and result files (in XML format) for the web application. The first section is usually hidden from the end user. The second section is the user and web interface of iHOP responsible for the displaying the result files produced by the iHOP factory and has stand alone capabilities. The gene/protein name dictionary of the system is assembled from public domain databases such as LocusLink and UniProt and was then extended to include spelling variants. iHOP also relies on indexing the abstracts for search and retrieval purposes. The system uses a relational database (PostgreSQL^{TM}) to store the text and gene data.

iHOP uses a custom created indexing procedure where the indexing process is split into two different steps. In the first process, hash code comparisons were used to search for parts of a gene synonym case sensitively. In the second step, taking the contextual information into account, the genes are allocated to their respective positions in the text. From the second step of the indexing procedure onwards, for every gene and abstract, an XML file is produced. The XML file produced contains the abstracts with genes, synonyms, MeSH terms and associated verbs marked and the gene document includes database references, homologous genes list and a list of synonyms. The database information is used by the web application to link to external databases.The iHOP factory transfers all the XML files and a small world wide web database to the web application. The web application uses this database to identify genes in the user query and to guide the user to corresponding XML file.

In iHOP query page (see Figure 3.9), the user is asked submit a query a gene name, accession number, NCBI gene entry or UniProt. A successful submission of query leads to the hit page, where the found entities are listed down with the organism and four icons each hyper linking to certain results. The first result is defining information for the gene, which gives the sentences in abstracts where the queried gene is defined. The second one gives interaction information of genes in the form of sentences form PubMed abstracts where the interaction information is found, the third link gives the most relevant information of the gene under study and the last one giving minimal information about the gene such as UniProt identifiers of the gene, OMIM identifiers, NCBI Gene, protein and nucleotide REFSEQ entries, Unigene id, Genbank accession and homologues of the genes in different organisms. iHOP also allows the creation of an interaction network aby adding interaction information to the existing network.

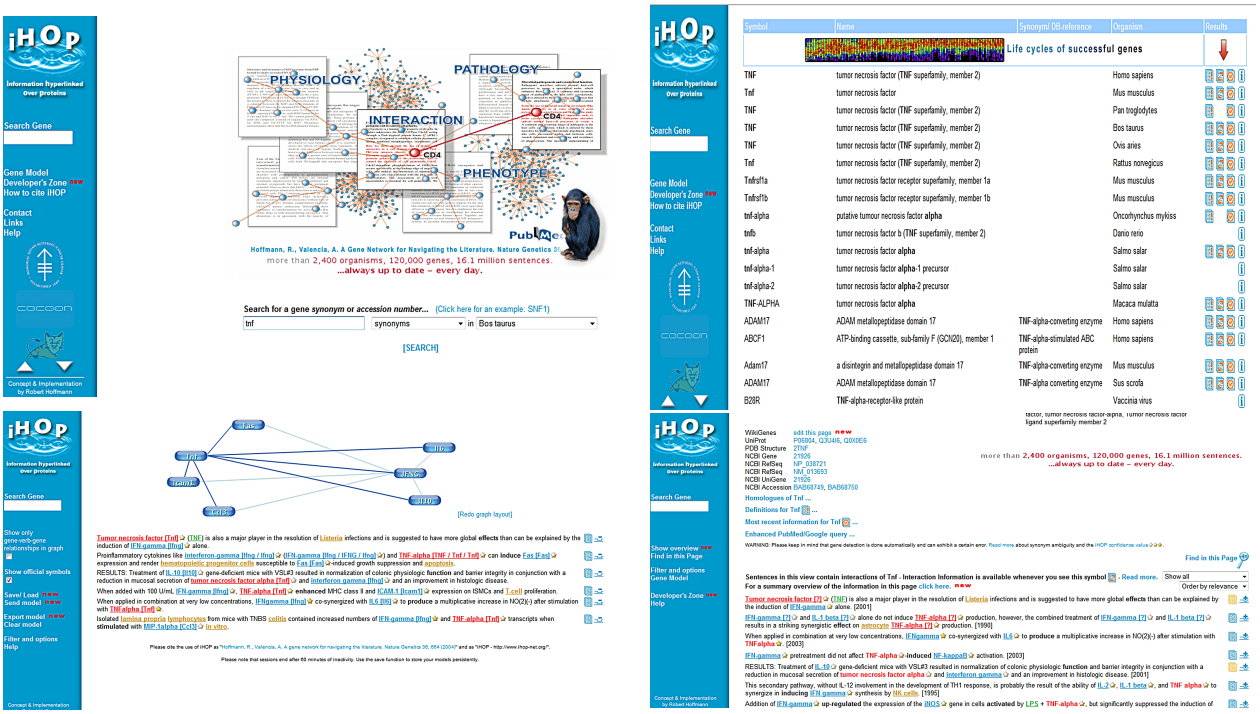


Figure 3.9: iHOP interaction pages (clockwise: iHOP home, iHOP search view, iHOP interaction results, iHOP network visualization)

Some of the commercial tools used in biomedical text mining and information retrieval include NextBio[8] and novoseek[9] by Bioalma[10].

The NextBio is an ontology based platform. The semantic framework of Nextbio is based on gene, tissues, disease and compound ontologies and contains documents from text based sources such as literature, clinical data and experimental data[11]. Figure 3.10 shows the result interface of NextBio.

[8]https://www.nextbio.com/b/nextbio.nb last accessed 18 November 2009
[9]http://www.novoseek.com/Welcome.action last accessed 18 November 2009
[10]http://www.almabioinfo.com last accessed 18 November 2009
[11]https://www.nextbio.com/b/corp/platform.nb last accessed 18 November 2009

Figure 3.10: Nextbio result page showing highlighted entities and popup explanations

Novoseek is a biomedical search engine for genes and proteins, chemicals, medical procedures, authors, body parts, tissues and subcellular components on PubMed records[12]. Figure 3.11 shows the result interface of Novoseek.

[12]http://lane.stanford.edu/howto/index.html?id=3956 last accessed 18 November 2009

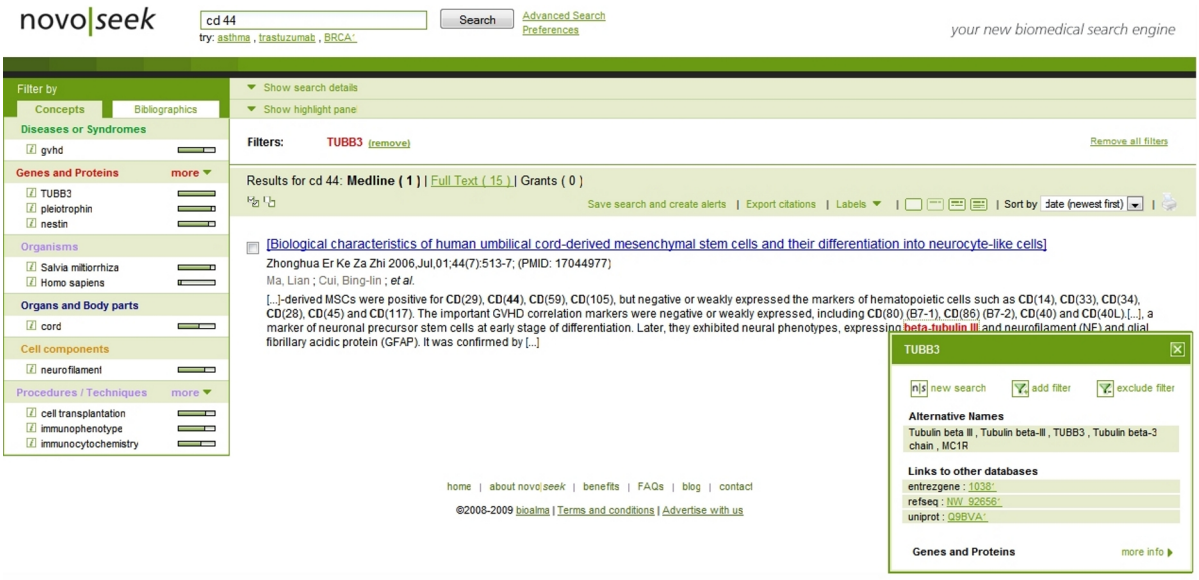


Figure 3.11: Novoseek result page showing external database information

Although the tools mentioned above work as potential NER systems, the initial requirements that these tools were designed to meet are different from the requirements that are under consideration here. It can also be seen that these tools do not function as knowledge representation systems but as NER systems capable of identifying gene and protein mentions in PubMed abstracts. Since these tools are designed to work for a large number of genomes, often the external database mappings provided are ambiguous and sometimes link to unrelated organisms. The extend of knowledge representation in these systems are also limited, often restricted to functional gene annotation from Gene Ontology and Entrez Gene or Uniprot database mappings. Most of the tools do not support ontology or terminology based search and retrieval. Finally, the above mentioned tools are restricted to recognition of gene and protein entities, where as SNP mentions and microRNAs, two other major factors of importance in genomics are left behind, which makes it necessary for a tool that is designed to meet with all the mentioned challenges and needs.

4 Ontology and Ontological search

Ontology is intended for the representation of the world (reality) or parts of it. An ontology describes a concept hierarchy that are related by the rules or order that defines the relationship in simple case, and in complex cases the relationships between the concepts are explained by principles and these principles are also used to restrict the possible interpretation of concepts (Guarino, 1998). Ontology provides a correct and consistent representation of entities relevant in a field, where as text mining identifies the collection of strings (words) that are used for communication as important. Current applications of ontology range from fields such as natural language representation to geographic information systems. Some instances of ontology from biomedical domain include GO (Gene Ontology) (Ashburner et al., 2000) functional gene annotation, Open Biomedical Ontologies[1] and UMLS (Unified Medical Language System) (Bodenreider, 2004).

Ontology developed in the biomedical domain can be a powerful tool for data integration and representation. Since researches in biological sciences are more knowledge based than hypothesis based, information retrieval and representation has more importance. Ontologies in the biological domain can be visualized by a Venn diagram of three domains, biological data, Computer science and Philosophy linguistics, biological ontologies can be represented at the area of intersection of the three domains (see Figure 4.1). In ontology concepts are defined as classes or groups of entities that belong to the same domain (gene, as a concept in genetics). Concepts are categorized into two different kinds, primitive concepts, those having just enough information to fall into a hierarchical division or a class. Defined concepts are concepts with sufficient information to be categorized into a specific class. Similarly, relationship between concepts can also be categorized. The first one is taxonomy, where concepts within the same classes are subdivided. The first division in taxonomy is specialisation relationship, and is commonly known as "is_a relationship" as the concept under consideration has all the features of the parent class, for example an enzyme *is_a* protein. The second one is partitive relationship, where concepts are defined as "part_of " or "hasComponent" relationship, for instance, enzyme *hasComponent* active site or active site is a *part_of* enzyme. The second kind of relationship is associative relationship. Associative relationship can be categorized into three, nominative relationship, which describes the names of concepts (enzyme *hasName* enzyme name), the second type of associative relationship describe the cellular location of one concept with respect to another (enzyme *hasCellularlocalization* cytoplasm). The third type of associative relationship describes the function or process that a concept is involved in (enzyme *hasFunction* hydrolase). Figure 4.2 illustrates the ontology building

[1]http://www.obofoundry.org last accessed 18 September 2009

life cycle.

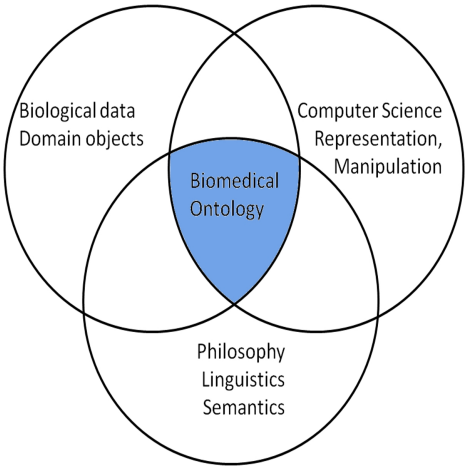


Figure 4.1: Venn diagram of fields intersecting in biomedical ontolgy adopted from Schulze-Kremer (2002)

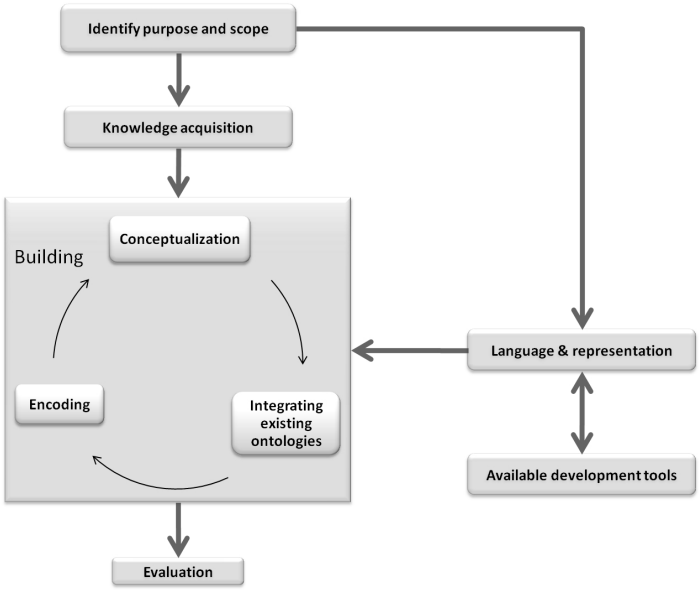


Figure 4.2: Ontology building life cycle adopted from Stevens et al. (2000)

Ontologies can be represented in the form of a graph, where each node of the graph represents a term or a concept and the edge connecting the nodes represent relationship between the connected nodes.Most of the ontologies are represented as a directed acyclic graph, where the edges of the graph are directional and there is no edges from the bottom layer of the hierarchy to the top layer. Ontologies as a whole can be classified into three major groups: Domain oriented ontologies, as the name suggests, ontologies with their scope limited to a particular domain (example, Mouse gross anatomy and

development ontology[2]). The second major group is task specific ontology, which are either specific for certain task or generalizations of certain tasks (example, Ontology for biomedical investigations[3] (Smith et al. 2007)). The final group of ontologies are the general ontologies that are for common high level concepts such as structure and substance. Currently, the ontologies developed in the biomedical domain are expressed in XML or knowledge representation language such as Resource Description Framework[4] (RDF) or Web Ontology Language[5] (OWL) or the OBO foundry format.

Three major fields were ontology developed in the biomedical domain can be of major importance are:

- **Data comparison**: Data in biological sciences is scattered throughout. It is not possible to figure out an all in one data source for biology. In the World Wide Web biological data is wide spread in different databases and science journals. What is really needed here is a comparison tool that could compare between different forms and kinds of data. Data comparison can be achieved if knowledge, concepts in data can be categorized into different domains and relations could be established between different domains concepts.

- **Data confinement**: The second field of importance for a biological ontology is data confinement and restriction. Biological databases hold large volume of data and biological data can be quite easily linked through underlying concepts. For example from genome information, going through underlying concepts between data one could end up in cytological or biochemical data, adding to it is the increasing amount of data being produced every day. Ontologies provide a hierarchical classification of different concepts and restricting the relations allowed for each domain facilitating concept localization (Stevens et al., 2000).

- **Data integration**: A third field where ontologies can be of help to biologists is data integration (Schulze-Kremer, 2002), where data from different domains in biology can be integrated by following the underlying concepts and relations which can be applied to relate the different fields together.

The major uses of ontologies in the biomedical domain include:- *Neutral authoring* (Community reference): All the different concepts in a domain can be expressed in an ontology, provided the ontology has sufficient coverage over the subject matter of interest. The benefits of such a representation include knowledge reusability, improved maintenance of the knowledge and long term knowledge (Uschold and Gruninger, 2004). *Defining a database schema or a common vocabulary for database annotation*: Using ontology as a specification can assure that a common vocabulary is available for knowledge description, sharing and querying. The advantages of using ontology as a database specification includes improved documentation, maintenance and reliability of the data and enable

[2]http://www.obofoundry.org/cgi-bin/detail.cgi?id=emap last accessed 18 September 2009
[3]http://www.obofoundry.org/cgi-bin/detail.cgi?id=obi last accessed 18 September 2009
[4]http://www.w3.org/RDF/ last accessed 18 September 2009
[5]http://www.w3.org/TR/owl-guide/ last accessed 18 September 2009

the reuse of knowledge. *Ontologies for common access to information*: The data in biological databases are accessed by different softwares that use different formats for data representation. This creates the need for format translators which can make inter conversion of different formats. The use of ontologies eliminates this problem by the use of a common agreed neutral format that can become the basis for format conversion and data mapping. Interoperability of format is an advantage of using ontologies for common access to information.

4.1 Major Biomedical ontologies

Ontologies and ontology development is in the main stream activity of bioinformatics. The use of ontologies in the biomedical domain started in the 1990's. One of the early biomedical ontologies is RiboWeb (Altman et al., 1999), developed to aid the construction of 3d models of ribosomal components and to compare the results with existing ones. RiboWeb is a collection of four ontologies, the physical-thing ontology which describes ribosomal components and associated structures as physical things, the data ontology which is to capture the knowledge in the experimental detail and data on the structural details of "physical things". The method ontology incorporate knowledge about the techniques applied to data and inputs and outputs of each method. EcoCyc (Encyclopedia of *Escherichia coli* K-12 Genes and Metabolism) is a database for genome information and biochemical machinery of *E. Coli*. EcoCyc uses ontology for database definition (Stevens et al., 2000).

The major wave of change in biomedical ontologies came with the implementation of Gene Ontology (GO)[6] (Ashburner et al., 2000). Gene ontology was developed to alleviate the problem of comparing the gene functional annotation of various organisms. The Gene Ontology project came into reality as the collective effort of three model organism databases, Flybase (Drosophila)[7](Gelbart et al., 1997), Saccharomyces Genome Database (SGD)[8](Cherry et al., 1998) and Mouse Genome Database (MGD)[9](Blake et al., 2003a). The GO project has developed three ontologies, Molecular function, Cellular component and Biological process. Molecular function accounts for catalytic or binding activities that occur at the molecular level. Cellular component describes the components of a cell or part of a large entity (anatomical structure or gene product group). Biological process explains the events that carried out by series of molecular functions[10]. Gene ontology is modelled as a DAG (Directed acyclic graph).

[6]http://www.geneontology.org last accessed 18 September 2009
[7]http://flybase.org last accessed 18 September 2009
[8]http://www.yeastgenome.org last accessed 18 September 2009
[9]http://www.informatics.jax.org last accessed 18 September 2009
[10]http://www.geneontology.org/GO.doc.shtml last accessed 18 September 2009

Table 4.1: Gene Ontology statistics as of 2 November 2009

28594 terms	99.0% with definitions
17454	biological_process
2482	cellular_component
8658	molecular_function

The relations modelled in Gene Ontology (see Figure 4.3) include **is_a** relation, **part_of** relation and **regulates** relation. In GO relations **is_a** relation is described as "is a subtype of" and does not include the relation "is an instance of". For two entities A and B the relation **part_of** in GO is stated only if B is a component of B and B exists only as a component of A and presence of B necessarily means presence of A and not vice versa. The relation **regulates** is modelled for two processes, if one process has a direct effect on the other. The two subtypes of the relation are **positively regulates** and **negatively regulates**. For two processes A and B the relation regulates can be modelled only if the process B regulates A whenever B is present.

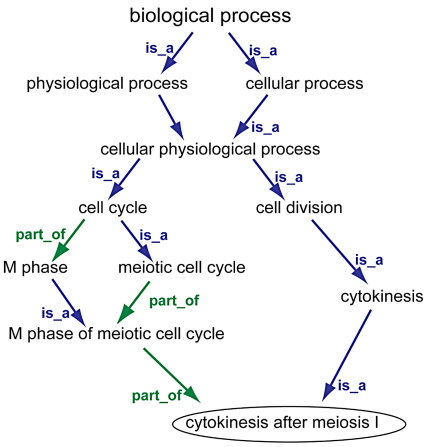


Figure 4.3: Relationships modelled in Gene Ontology Relationships modelled: **is_a**: is a subtype of, **part_of**.
adopted from http://www.yeastgenome.org/help/images/cytokinesisDAGrels.jpg
last accessed 18 September 2009

The success of Gene Ontology saw the development of similar ontologies in the biomedical domain for different purposed. The OBO consortium[11] (Smith et al., 2007) was formed for coordinating these efforts. The current principles of OBO foundry are[12]:

[11]http://www.obofoundry.org last accessed 18 September 2009
[12]http://www.obofoundry.org/crit.shtml last accessed 18 September 2009

- The ontology must be open and available to use for all, the origin must be acknowledged and cannot be altered and redistributed under original same or with original identifiers.
- A common shared syntax can be used for expressing ontology; it can be OBO syntax, its variations or OWL (Golbreich et al., 2007).
- There is a identifier space for each ontology within the OBO foundry.
- The content of the ontology is clearly specified and described.
- There are textual definitions for all terms in the ontology.
- The relations used in ontology are unambiguously defined in OBO relation ontology (Smith et al., 2005) and the ontology is well documented.
- The ontology has a large group of independent users.
- And OBO members collaboratively develops the ontology.

OBO foundry has candidate ontologies in domains ranging from anatomy to taxonomy (see Figure 4.4).

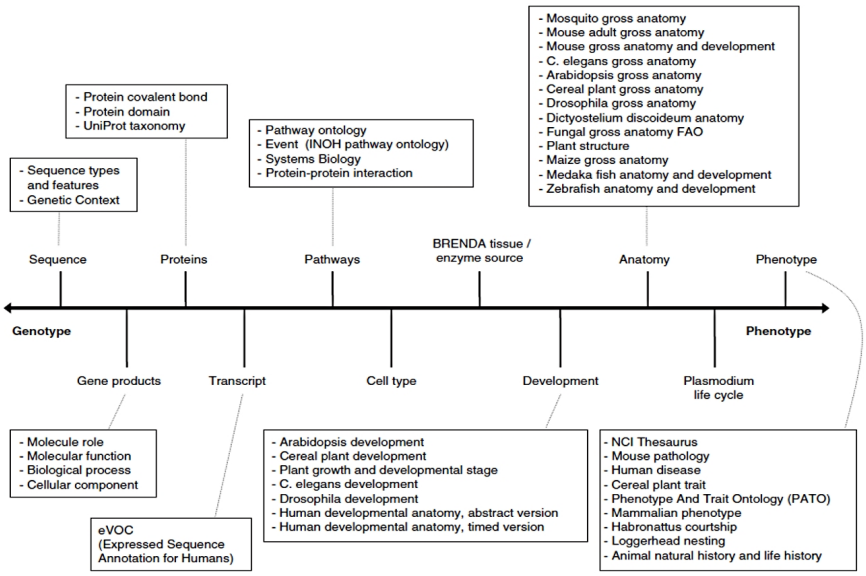


Figure 4.4: OBO ontologies arranged on a spectrum adopted from Bodenreider and Stevens (2006)

Two of the most important ontologies in the OBO foundry are Gene Ontology and Sequence Ontology(SO) [13](Eilbeck et al., 2005). The former is explained in the beginning

[13]http://www.sequenceontology.orglast accessed 18 September 2009

of this section. The latter is an ontology for genomic annotation, designed to provide analysis, management and exchange of genomic data.

The SO project aims to describe the features and properties of biological sequences. The SO project is jointly undertaken by Flybase, Wormbase[14], Mouse Genome Informatics (MGD) and Sanger Insitute [15]. The ontology is based on features such as gene, exon, promoter and binding site that can be located on a sequence with coordinates. These properties of sequences are used to describe the attributes of the feature, such as sequence attributes (Maternally_imprinted_gene), consequences of mutation (mutation_affecting_editing) and chromosome variation (aneuploid). Sequence ontology is modelled as a DAG (Directed acyclic graph). There are three basic relations modelled in Sequence Ontology: **kind_of**, **derives_from** and **part_of** (Eilbeck et al., 2005) (see Figure 4.5). The relationship **kind_of** is modelled like the relationship **is_a** in GO and describes a subset relationship. Similar is the case with **part_of** relationship, which can be related to **part_of** relationship in GO with part whole relationship. The relationship **derives_from** describes relationship between two entities A and B, where the entity B is derived from A and existence of the entity B depends on the existence of A.

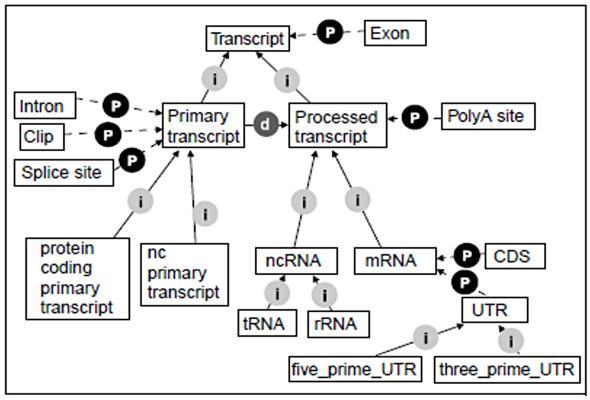


Figure 4.5: SO relations adopted from Eilbeck et al. (2005)
Legend: **d**:**derives_from**, **i**:**kind_of**, **P**:**Part_of**

4.2 Ontology based search

Ontology based search is a concept that developed in the late 1990's. One of the first uses of ontology based search was for personalizing World Wide Web search (Pretschner and Gauch, 1999). Ontologies can be used as structuring criteria for information repositories. This structuring enables better organization and classification of database repositories and repository indexing to be done based on this ontology (Uschold and Gruninger, 2004). For ontology based annotation of the repository, the documents in the repository should be annotated with the terms in ontology or mapping

[14]http://www.wormbase.org last accessed 18 September 2009
[15]http://www.sanger.ac.uk last accessed 18 September 2009

of the relevant terms in the document towards ontology should be done. This method of mapping to the terms in ontology and indexing the document with terms in ontology can help in the search and retrieval of documents. As the ontologies can be represented as a graph, each node of the graph holds a set of documents in which the term representing the node is the most "frequently appearing key term". For a query made using the key ontological terms, document retrieval is based on the key terms in the query and further refinement is done by following relationships in ontology. Ontology based document search in biological sciences has a centre stage in retrieving documents of relevance and also documents in the associated fields since associative information has greater importance in this field. For example, a search query 'apoptosis proteins' in AmiGO gene ontology browser [16] has returned results for 'induction of apoptosis', a GO biological process term. Along with this search result, results like 'positive regulation of apoptosis' and 'regulation of apoptosis' were found. These results are returned by the search engine, as a result of relationship modelling and associated search where all the protein data from associated nodes for apoptosis, especially from those nodes with an "is_a" relationship with node is returned. Some instances of ontology based search in biomedical realm include Gene Ontology Annotation (GOA) database [17] (Camon et al., 2004) where all the UniProt protein entries are annotated using Gene Ontology and GOPubMed [18] (Doms and Schroeder, 2005) where PubMed can be searched using Gene Ontology.

4.2.1 Semantic search

Semantic search used semantics of the keywords to produce relevant search results. There are two major kinds of searches performed by the user, navigational searches and research searches (Guha et al., 2003). Navigational searches are defined as searches were the user query is a phrase or combination of words. In these searches the user is using the search engine as a navigational tool to a predefined document. In this case semantic searches cannot be applied. The second kind of searches are research searches, where the user is trying to gather a number of documents with a search query, where semantic searches can be applied. The search queries for a research search denotes real world concepts, and search engines use semantics (study of the meaning of words) to understand the concept behind the search terms. The results obtained through these searches are added to traditional search results to produce diverse results. Semantic search has special importance in the search and retrieval of biomedical documents since documents containing information from different biomedical domains are interlinked by concepts and meanings. Semantic similarity searches are achieved by analysing the topological similarity between two documents or terms using ontology. Semantic similarity between two sets of documents or terms under consideration is determined by the concepts represented by them and the degree of relatedness between the concepts, determined by the semantic relationship between concepts. Certain statistical methods are also used to

[16] http://amigo.geneontology.org/cgi-bin/amigo/go.cgi last accessed 18 September 2009
[17] http://www.ebi.ac.uk/GOA last accessed 18 September 2009
[18] http://www.gopubmed.org last accessed 18 September 2009

determine semantic similarity. In the biomedical domain, primarily ontologies (especially Gene Ontology) are used to establish semantic similarity.

Semantic similarity searches are used for wide variety of applications in biomedical domain. Semantic similarity methods are used to validate automated annotation of gene products and result validation (Pesquita et al., 2009). Several prediction systems, for predicting protein-protein interaction networks (Zhu et al., 2007) and structural similarity of protein surface (Liu et al., 2007) use semantic similarity search method. It has been also used in protein clustering validating automatic protein annotation
(Couto et al., 2006). Gene Ontology is the base ontology for many of the semantic similarity comparison tools in the biomedical domain. FuSSiMeG[19] compares semantic similarity between gene ontology annotations of gene products in order to establish functional similarity between them (Couto et al., 2003). GOToolbox[20]
(Martin et al., 2004) is a collection of tools that used Gene Ontology resource for the clustering of functionally related genes within a data set, retrieval of genes sharing annotations with a query gene and identification of statistically under or over represented terms in a gene data set. G-SESAME[21] (Du et al., 2009) is an online tool for measuring functional similarities of gene products and semantic similarities of Gene Ontology terms. Several search engines in biomedical domain also incorporate semantic search capability. GoWeb[22] is a life science semantic search engine which integrates key word search with ontologies and text mining technologies. Hakia PubMed[23] is an ontology based semantic search engine. For retrieving information on health Hakia uses 10 million PubMed abstracts.

[19]http://xldb.fc.ul.pt/rebil/ssm/ last accessed 24 September 2009
[20]http://burgundy.cmmt.ubc.ca/GOToolBox/ last accessed 24 September 2009
[21]http://bioinformatics.clemson.edu/G-SESAME last accessed 24 September 2009
[22]http://gopubmed.org/web/goweb last accessed 24 September 2009
[23]http://pubmed.hakia.com last accessed 24 September 2009

5 Problem definition and goals

The problem definition comes from the animal science/animal genomics field, where scientists at the Institute of Animal science Animal husbandry/ animal genetics group, University of Bonn are in need of a tool for mining gene, protein, associated networks and related information concerning cattle preimplantation period, and meat quality traits in pigs. The specific needs of the researchers are concentrated on two specific fields. The first one is cattle preimplantation genetics, where current focus is on topics like genetics involved in developmental competence of preimplantation embryo from in vitro and in vivo sources, influence of various gene manipulation techniques on gene expression and developmental competence of in vitro and NT embryos. The second topic of research importance is pig genetics where much attention is given to on candidate genes responsible for meat and carcass quality and disease resistance.

There are several online tools available, which could identify genes proteins and interaction networks from scientific publications in biology (refer Section 3.6). But all these tools mentioned focus on a large number of genomes. Although focusing on a large number of genomes (organisms) enables a wide coverage for the tools, the specificity is lost. The wide coverage of the tool can also lead to noise in the final result. In most of the tools data enrichment is done only by mentioning the functional gene annotations and protein pathways. The rest of the information content in the biological databases are left behind, especially gene protein synonyms and sequence information from Entrez Gene and UniProt, SNPs associated with the gene from dbSNP and protein domain and family information from InterPro. This point to the fact that the user gets only a part of the knowledge and some vital information is left behind. Another factor is the method of document retrieval and sorting for result display. Majority of tools use PubMed as the base search engine, and the documents retrieved entirely depends on the expertise of the user in forming a search query as neither semantic search nor ontology based search and retrieval of the document are implemented. As a result a number of documents and information content in them are left out since those documents treat the key terms in the query as implicit concepts and not as explicit mentions. So the user may not be able to observe the whole picture of the domain that is searched for, instead will be restricted to a short collection of documents where the search query terms are found. A tool that incorporates all the above mentioned functionalities would be more apt to the requirements of the researchers in the domain. ProMiner, along with the knowledge discovery suite SCAIView is being used as a knowledge discovery platform in human mouse and Arabidopsis genomics and related domain and could be adapted to work in different fields according to the needs of the end user. The system incorporates ProMiner as a text mining tool and the results are displayed using SCAIView knowledge environment. ProMiner results displayed in SCAIView are enriched with

external database information and provides document sorting based on the number of entities found in the document and relevance of entity. The system also incorporates an ontology based search.

The current version of SCAIView, which is adapted to work with human and mouse genes and proteins can work cattle and pig data since all mammals have more than ∼80% gene sequence similarity among them, on the other hand, information on genes and proteins that are unique to cattle and pigs will be left behind and external database mappings in the present version are specifically tailored to fit human mouse and Arabidopsis genes. Hence, the system needs to be adapted to suit the needs of researchers. To utilize the full functionality of SCAIView in the livestock genomics field, the system needs an ontology or a terminology that could represent the key domains in livestock genomics. As of now, a specific knowledge discovery tool dedicated to livestock genomics has not been developed. So, the livestock genomics version of SCAIView developed as a part of this thesis project is intended as a prototype system for the introduction of knowledge discovery methods and text mining tools to the livestock genomics field.

The final goals of this thesis include:

- Introduction of text mining tools specifically devoted to animal genetics.
- Generation and curation of gene and protein name dictionaries for cattle and pig. Generation of microRNA dictionaries for cattle pig and the model organisms (human and mouse). Adapting SCAIView include generation of cattle and pig gene and protein name dictionaries, as ProMiner is a dictionary based system. The generated dictionaries need to be curated to remove ambiguous entries. Gene/protein name entries from the final result should be mapped to external databases and Gene Ontology functional gene annotations for integrating external database information to SCAIView along with ProMiner results.
- Performance evaluation of animal science version of SCAIView.
- Terminology analysis concerning cattle preimplantation period. Condensing the entire information about cattle preimplantation period into a set of key terms representing key concepts in cattle preimplantation domain in order to aid the search and retrieval of documents.
- Analyzing the network of interacting gene products in cattle preimplantation period. Analysis and validation of protein-protein interaction network (from co-occurrence) retrieved from SCAIView validation of interaction networks obtained from SCAIView and suggesting potential novel interaction partners.

6 Background work: use of bioinformatics tools and techniques in livestock genomics

The past decade has been a time of rapid advancement for livestock genomics, during the past years the livestock genomics has grown from synteny maps into complete genome sequences and molecular level functional descriptions. Additionally, expression profiling of animal genes are also being done, adding to the already existing vast amount of data that is present in the farm animal genomic field. With the time it became a necessity that farm animal genomics should follow the footsteps of human and murine genomics in digital data handling, storage and retrieval. Although farm animal genomics has been lead by human genomics, animal scientists were able to adapt the existing tools for their needs. The use of bioinformatics tools and computational techniques in animal genomics can be categorized into two major sections:

- Farm animal genome databases (the use of computational data storage).
- The use of bioinformatics tools (data retrieval, visualization and manipulation-tools).

6.1 Farm animal genome databases

Farm animal genomics has dedicated genome databases for partially and completely sequenced genome projects. But, as the scope of this thesis is limited to cattle and pig a brief survey of some of the existing cattle and pig genome databases was done as a part of this project and is given below:

6.1.1 Survey of existing cattle genome databases

Cattle genome coordination program

The cattle genome co ordination program is one of the projects supported by the National Animal Genome Research Program (NAGRP) and comes under U.S Livestock Species Genome projects. The program does not directly host a database, but provides links to all the existing databases. NCBI/genbank, Ensembl genome browser, Cattle genome information at NCBI, Bovine Genome Database are some of the databases to which the cattle genome co ordination program provides direct links. It provides direct links to

cattle genome maps, QTL maps, genome database maps and fingerprint maps that are hosted by different databases.

Bovmap Database

The INRA Bovmap Database[1] (Law and Archibald, 2000) is funded by the French National Institute for Agricultural research (INRA). Information available from Bovmap includes chromosome map assignments of genomes, polymorphism details, sequence homologies and PCR primer sequences.

The search and retrieve options in the database include: a chromosome map (mapped on the basis of loci information) with links to all the loci in the chromosome. The respective locus page has all the basic information including locus name, locus type, synonyms, classification, gene family, homology with other species, and effect of the locus and Genbank/EMBL reference. This page also provides cross reference to Genbank, for sequence retrieval, links to cattle chromosome physical and genetic map and bibliographic reference to pubmed. Bovmap database provides links to Gene atlas[2] and Gene cards[3] databases also. The other search and retrieval tools in the database include: retrieval of polymorphic locus information based on different breeds, locus information, homology information, QTLs, Gene lists, sequence retrieval and chromosome homology. The database does not provide adequate web links to other databases and some of the provided links are dead.

Table 6.1: BovMap Data statistics as of 1 June 2009

LOCI: 4357	ASSIGNED LOCI: 4125
GENES: 1558	ASSIGNED GENES: 1507
MICROSATELLITES: 2402	ASSIGNED MICROSATELLITES: 2244
POLYMORPHISMS: 3100	ALLELES: 551
816 BREED POLYMORPHISM records on 103 breeds	
PROBES: 169	PRIMERS: 4764
ENZYMES: 2733	
6240 HOMOLOGUE LOCI on 156 SPECIES	4697 HOMOLOGUE LOCI LINKS on 67 SPECIES(DBs)
BIBLIOGRAPHIC REFERENCES: 966	
GB EMBL REFERENCES: 0	SWISSPROT REFERENCES: 1019
47 EXTERNAL DATABASES (28074 active links)	

[1]http://locus.jouy.inra.fr/cgi-bin/bovmap/intro.pl last accessed 14 June 2009
[2]http://www.geneatlas.org last accessed 14 October 2009
[3]http://www.genecards.org last accessed 14 October 2009

The Bovine Genome Database

The Bovine Genome Database[4] is run by Bovine Genome Sequencing and Analysis Consortium. The consortium includes members from Georgetown University, University of Minnesota and CSIRO (Commonwealth Scientific and Industrial Research Organization) Livestock Industries. The bovine genome database has the bovine genome sequences that are downloaded from genbank and clustered using ENSEMBL protein dataset (version 33). The database services provided are: QTL viewer, which provides a list of QTLs and individual QTL pages provide information such as QTL Id, trait, appearing chromosome, family, lod score and references for the QTL. This page also provides a map of the respective chromosome with position of QTL on the chromosome. Bovine Genome Database also provides information on ESTs (Expressed Sequence Tags), where the inputs can be accession identifiers from a variety of databases like Genbank, ENSEMBL, Swiss-Prot, RefSeq[5] and SPTREMBL (SWISS-PROT TrEMBL). The search results provided include a list of all the possible hits with links to Ensembl, UniProt and NCBI protein database. The result page also provides GO annotation of the ESTs. The other searches provided include chromosome search to extract all the information on a specific chromosome, and scaffold search to retrieve information on cattle genome scaffolds. The database provides visualization of specific regions of a chromosome through composite map and BAC[6] map. This database also has a re-PCR page (reverse electronic PCR), which allows the user to identify SNPs and QTL regions. The re-PCR outputs are printed as tables containing the chromosome position and QTL regions in the output table, and provides link to the ENSEMBL database.

Bovine Genome database also allows the user to BLAST sequences against *Bos taurus* genome. The BLAST interface has three BLAST options, MegaBLAST, BLASTn and BLAST oligomeric sequences. Each time a BLAST is done, the database allows the user to retrieve SNPs that are present in the region identified. The database provides all the basic information needed by the user but do not include a keyword search for fuzzy searches.

Cattle Genome Database (CGD)

The Cattle Genome Database[7] is hosted at the Queensland Biosciences Precinct. The CGD is a part of an international collaboration to map the Bovine genome. CGD includes: Chromosome maps for all the 29 autosomal chromosomes and X and Y chromosomes. Comparative map, with the locations of genes mapped by linkage analysis in cattle and comparison to pigs, rats mice and human genome. The search tools provided by database includes: Locus search, pair wise Lod score of loci and search for QTLs (Quantitative Trait Loci). The chromosome maps provided by the database are graphic images and do not provide an interactive visual interface. The search tools used in the database provide

[4]http://genomes.arc.georgetown.edu/bovine last accessed 14 June 2009
[5]http://www.ncbi.nlm.nih.gov/RefSeq last accessed 14 June 2009
[6]Bacterial Artificial Chromosome used in sequencing projects to amplify target organism DNA
[7]http://cgd.csiro.au last accessed 17 June 2009

with the minimal information and do not provide a mining/sequence retrieval interface.

6.1.2 Survey of existing pig genome databases

NAGRP Pig genome coordination program

Supported by the National Animal Genome Research Program (NAGRP), the coordination program provides direct link to various databases that are supported by various organizations. Some of the database links in the web page include: pig genome at pre-Ensembl (unfinished), NCBI Pig genome project, Danish-Chinese pig genome project and Japan Pig EST Data Explorer.

Pig Genomic Informatics System (Pig Analysis Database)

Pig Genomic Informatics System[8] (PGIS)(Ruan et al., 2007) is an online database of pig genome and is designed to identify pig genes based on homologous human genes and SNPs. The Pig Analysis Database integrates 3.84 million whole genome shot gun reads, 0.7 million ESTs (Expressed Sequence Tags) generated by Sino-Danish Pig Genome Project and 1 million Genbank records. To identify pig genes, the pig genomic sequences were aligned with human genomic sequences using BLASTX, from the results so obtained, the regions with low identity were discarded. As the next step repeating sequences were masked, and the resultant sequences were aligned to human exons by cross matching. To identify SNPs, high quality genomic sequences of five pig strains in WGS reads and ESTs were compared and the variations in base pairs were identified as SNPs. PigGIS can be browsed to fetch information by clicking on human genome displayed on the home page or information can be queried using a keyword for search. A keyword search with a gene -symbol will give a result page with the respective ENSEMBL gene and transcript ids, location in the chromosome, description and sequence information. The other search options in the database include:

BLAST search:-to look for similar sequences in the databases.

Oligo design: - to design an oligo nucleotide based on the given sequence information.

The database incorporates some visualization interfaces, which include:

- Transcript view-shows consensus of the pig coding sequences
- Gene view- presents various splicing forms,
- Exon view- alternative sequences given in a human exon,
- Cluster view-assembly of contigs,
- SNP view-details about pig SNP and Sequence view which present complete information of raw sequences.

[8]http://pig.genomics.org.cn last accessed 14 June 2009

The database contains valid information, but the amount of information is less, does not allow the user to mine and retrieve data.

Table 6.2: PGIS database statistics as on 1 June 2009

Reads	Hampshire	707,281
	Yorkshire	1,204,666
	Landrace	650,609
	Duroc	1,015,722
	ErHuaLian	256,993
	Total	3,835,271
EST		870,084
Homolog Gene (Human vs. Pig)	Reads covered	14,618(66%)
	EST covered	12,299(55%)
	Covered by both	16,309(73%)
	+Genbank seq	16,958(76%)
Oligo		20,789
SNP		16,965

Swine Genome Mapping Project

The Swine Genome Mapping project is funded by the United States Department of Agriculture, Agriculture Research Service. The services provided by the database include: an interactive linkage map, which visualizes all the markers that are present in each chromosome and allows the user to access the marker information. The marker information page includes information such as chromosome, locus, marker type, the submitted database, other information and a list of references to the marker. The marker table provides information about markers by the respective chromosome and relative position. The database also allows the user to search for marker information by name, and other data provided include clones identified for each locus and overgo oligos (oligo nucleotides conserved across species). The database does not provide direct links to cross referenced databases.

Pig Expression Data Explorer (PEDE)

Pig Expression Data Explorer[9] is a database of full-length cDNA clones and ESTs in pigs, maintained by the Animal Genome Research Program, Japan and conducted by National Institute of Agro biological Sciences and STAFF-Institute. The database was put together from sequences that are assembled from porcine 5' EST sequences. The search and retrieval tools in the database include: the cluster viewer page (in the old version of the database), the interface which allows searches by keyword, locus name and database accession number. The database allows BLAST-ing on clone or assembly sequences, and retrieving the sequences data of interest. The BLAST result page is provided with links to the obtained sequence assemblies/clones. The sequence assembly/clone information

[9]http://pede.dna.affrc.go.jp last accessed 14 June 2009

page gives links to similar assemblies/clones, to the related human genome, and links to RefSeq and UniGene databases. This page shows other information such as nucleotide and amino acid sequences, includes a link to retrieve the sequences in FASTA format, shows similarity of the sequences to human genome and displays similar proteins or peptide chains that are identified in different databases through BLAST-ing the sequence. The GO viewer interface classifies clones or assemblies according to Gene Ontology terms. PEDE also includes a Pig cSNP (SNPs in cDNA) database, identified from ESTs assemblies of PEDE; the result view provides the user with SNPs that are present in some of the pig breeds and an estimate of the amount of sequences that are needed to encode the full-length CDS (Uenishi et al., 2004). The website also has the option to download all the clone/assembly sequences. The database does not have a browse and fetch interface for the users to browse through genomes and to retrieve the data needed and some of the links provided by the data base are found to be dead.

Animal Genome Research Program (Swine)

Animal Genome Research Program (AGP) is a joint project of National Institute of Agrobiological Sciences (NIAS) and Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries (STAFF-Institute) Japan. The database currently includes only Swine genome and the services provided by the database include: Swine linkage map viewer, plotting linkage distance between loci on a chromosome. The Swine marker viewer gives information about marker loci, which has the information about the forward and reverse primer of the marker and linkage analysis information along with RH (Radiation Hybrid)[10] information. All together, the database gives only the locus centered information and gene positional information.

Recent developments in animal genomics databases include, ANEXdb , animal expression database which allows the user to store and analyze gene expression data in the form of microarray experiments. ANEXdb[11] is a joint venture by the Iowa state university, National Science Foundation (NSF) and United States Department of Agriculture (USDA). AgBase[12] (McCarthy et al., 2006) is a curated web accessible and open source resource for functional analysis of agriculture plant and animal gene products. AgBase is maintained by the Missisipi state university. AgBase provides the user with tools for the functional annotation of genes with GO terms.

[10]Chromosomes are separated from one another and high dose X rays are used to break into several fragments. The order of markers on a chromosome can be determined by estimating frequency of breakage. RH mapping is used to create whole genome radiation hybrid map

[11]http://www.anexdb.org last accessed 7 September 2009

[12]http://www.agbase.msstate.edu last accessed 7 September 2009

6.2 MicroRNA database

miRBase[13] is the major repository and resource for microRNAs data now. Each entry in the database is a predicted hairpin portion of a microRNAs transcript and is termed mir in the database, 'miR' is the term given to mature miRNA sequence in the database with sequence and location information. The database is currently hosted and maintained by the faculty of Life Sciences at University of Manchester and is supported by Biotechnology and Biological Sciences Research Council (BBSRC), United Kingdom. The database has four major divisions:

miRBase Registry: The registry service provides the users with unique microRNA names prior to publication in peer reviewed journals (Griffiths-Jones et al. 2006).

miRBase Sequence: miRBase sequence provides the users with sequence and annotation data and references and links for microRNA data (Griffiths-Jones et al. 2006).

miRBase Targets: This service provides the user with an automated pipeline for the prediction of targets for all animal miRNAs that have published data. The service uses miRanda algorithm for the prediction microRNA targets (Enright et al. 2003). Recently, mi Base Targets has been renamed as microCosm[14] and is now hosted at EBI.

miRBase Genomics: miRBase Genomics gives information on genomic regions surrounding miRNA precursors, such as transcription start sites, CpG islands, ESTs, cDNAs, genomic repeats, conserved transcription factor binding sites, expression ditags and polyA signals[15]. Currently the scope of miRBase genomics is limited to genome of *H. sapiens*, *M. musculus*, *R. norvegicus*, *C. Elegans* and *D. Melanogaster*.

The database has also introduced a nomenclature schema for microRNAs. Certain key components of the nomenclature schema are:

- The miRNA name has a three or four letter abbreviation of the species name as prefix and a numeric suffix.
- If more than one hairpin precursor expresses the same mature miRNA, each of them are denoted with numeric suffixes (example, dme-mir-6-1 and dme-mir-6-2)
- Lettered suffixes are given to related hair pin loci that expresses related mature miRNA sequences (example mmu-mir-181a and mmu-mir-181b).
- For plant miRNAs different nomenclature schema is used, plant miRNAs are of the form ath-MIR166a. The suffixes describe distinctive loci which express all related miRNAs and numeric suffixes are avoided.
- Conventionally viral miRNAs are named after the locus form which they are derived.

The current version of the database is release 14 and it contains 10883 hairpin precursor miRNAs entries and 10581 mature miRNA products, in 115 species. In the new version, 1367 new hairpin sequences and 1580 novel mature miR and miR* products have been included along with the previous data.

[13]http://www.mirbase.org last accessed 18 September 2009
[14]http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5 last accessed 10 November 2009
[15]http://www.mirbase.org/genomics.shtml last accessed 5 November 2009

6.3 Use of Bioinformatics tools and softwares

The bioinformatics tools used in animal genomics can be categorized into six major sections such as:

- Sequence analysis software,
- Genetic analysis software,
- Functional Genomics Software,
- Genetic networks and synthetic biology,
- General Application software and
- other comprehensive tools[16].

The tools are further classified according to the different uses in the major classification section. For example, softwares used in sequence analysis is further classified as tools for base calling, assembling, viewing, SNP analysis, RNAi searches and so on (Table 6.3).

In genomics, statistical and data analysis tools are used in microarray and gene expression data in two different directions. The first one includes data normalization and calibration and the second one is the statistical analysis itself. In livestock genomics, the major statistical tool used is 'R' and the analysis packages that are available in the R homepage[17] and website of the Bioconductor[18] (Gentleman et al., 2004). The popular data normalization 'R' packages include 'marray' and 'vsn'.

[16]http://www.animalgenome.org/bioinfo/resources/intro.html 14 September 2009
[17]http://www.r-project.org last accessed 19 October 2009
[18]http://www.bioconductor.org project last accessed 19 October 2009

Table 6.3: Bioinformatics tools in livestock genomics

Major animal genomics fields	**Bioinformatics tools used**
Sequence analysis	
Base calling, assembling, viewing	Phred, Phrap, Consed
SNP Analysis	Consed, DnaSP, PolyPhred, SNPidentifier
Alignment / Clustering	Arachne, Assembler, CAP3, ClustalW
RNAi searches	miRanda, MiRFinder
Similarity Search	Blast, Blastz, Fasta, GMAP, Mulan
Primer design	Amplify, CodeHOP, Expeditor
Genome Analysis	Light weighted genome viewer, Vmatch
Genetics Analysis	
Linkage analysis	CarthaGene, Crimap, JoinMap
Haplotype analysis	fastPHASE, PHASE
Genetics analysis	epiSNP, EPISNPmpi, MiNiInbred
QTL analysis	BmapQTL, MapQTL, QTL Cartographer
Phylogeny analysis	Phylip, PAUP, PhyML, TimeTree
Pedigree	Pedigraph, Pedigree-Draw, MiniInbred
Functional Genomics Software	
Domain/Motif Prediction	EST Scan, ORF Finder, ORFpredictor
Gene Prediction	Augusrus, HMMgene, GENEscan, Grail
Microarray Analysis	ArrayGenes, KegArray, HDB Stat
Genome analysis	GenomatiX, PepTool, Genewiz
Annotation and Gene Ontology	Apollo, AmiGO, ENSMART, GoSurfer
Expressoin data analysis	DAVID and EASE, GenMAPP, Gaggle
Comparative analysis	COGs, PEDANT 3, KEGG, TheSEED
Genetic networks and synthetic biology	SynBioSS

7 Materials and methods

This chapter describes the materials and various methodologies that are used. First section, Materials includes descriptions about various databases used for retrieving the data and the tools used for data generation, analysis and manipulation. Second section describes the methodologies followed.

7.1 Materials

The different databases and tools that were made use of in this thesis are explained in this section.

7.1.1 Entrez Gene

The gene specific database by National Center for Biotechnology Information (NCBI) and one of the databases in Entrez family of databases. Entrez Gene[1]
(Maglott et al., 2005) contains genes from the genomes that are sequenced and do not include predicted genes. The data taken form NCBI Reference Sequence project (RefSeq), other NCBI databases and collaborating databases, curated and automatically integrated to Entrez Gene. The data content includes nomenclature, map location, gene products and their attributes, markers, phenotypes, and links to citations, sequences, variation details, maps, expression, homologs and protein domains and is updated as soon as new data becomes available.

Entrez Gene was used to retrieve cattle and pig gene synonyms for dictionaries and for external database mappings, Entrez Gene is used as the key identifier to which most of the external database identifiers are mapped. Entrez Gene also provides a list of PMIDs as references for each entry Entrez Gene. This was used to create a reference list of Pubmed documents mentioning cattle and pig gene (gene2pubmed file from Entrez Gene ftp service). The cattle genome build 4.1 data and pig genome build data 1.1 was used in this thesis.

7.1.2 Ensembl

Ensembl[2] genome browser database which is a joint project between EMBL-EBI and Wellcome Trust Sanger Institute. The database contains data from sequencing projects undertaken by the Wellcome Trust Sanger Institute. It is updated every six months and

[1]http://www.ncbi.nlm.nih.gov/gene/ last accessed 1 December 2009
[2]http://www.ensembl.org/index.html last accessed 1 December 2009

the current release is Release 55, which contains data from fully sequenced genomes of 47 species and chromoseome assembly from 3 partially sequenced genomes. Ensembl include gene and protein data such as gene and protein names and a few synonyms, sequence information, chromosome location of the gene, various splice variants of the gene, supporting experimental data, external database references and data on regulatory genes for a gene.

Various services provided by the Ensembl database are used in this thesis. Ensembl database has integrated BioMart (Smedley et al. 2009) data retrieval system into its services. The BioMart service provided in the Ensembl database was used for data retrieval purposes. The service was used for orthologue mapping of cattle genes to the corresponding human and mouse genes and external database mapping for cattle genes. The database provided mappings towards external databases like UniProt and InterPro and mapping Ensembl identifiers towards Entrez Gene identifiers.

7.1.3 UniProt (Universal Protein Resource)

UniProt[3] (Bairoch et al., 2005) is collaboration between EBI, SwissProt and PIR and provides a single, centralized source for protein sequence and functional information. Uniprot consortium was formed by uniting Swiss-Prot, TrEMBL and PIR protein sources. UniProt has three different component databases, UniProt Archive (UniParc), UniProt Knowledgebase (UniProtKB) and UniProt Reference Clusters (UniRef). UniParc provides non redundant sequence data, UniProtKB is the central database with protein sequences and functional information, UniRef has clusters of UniProtB proteins based on sequence similarity.

UniProt data was used in the generation of gene and protein name dictionaries for cattle and pig from cattle and pig gene/protein annotations (for creation of curated dictionaries). The data from UniProt mapped to UniProt identifiers were used to map Gene Ontology identifiers to UniProt data for data enrichment. Like Entrez Gene, Uniprot also provides a list of PMIDs to identify PubMed documents from which the protein information was retrieved. The PMIDs mapped to UniProt identifiers were retrieved and used for generating a reference list of PMIDs and was later merged with reference list from Entrez Gene. For this work the Uniprot release version used was 15.3 released on 26 May 2009.

7.1.4 KEGG (Kyoto Encyclopedia of Genes and Genomes)

A database for gene functions and linking gene functions to higher order functional information. KEGG[4] (Kanehisa and Goto, 2000) was initiated in 1995 by Japanese human research program. KEGG is a collection of five databases namely, KEGG Atlas, KEGG Pathway, KEGG Genes, KEGG Ligand and KEGG BRITE. The data in KEGG databases are updated regularly.

[3]http://www.uniprot.org last accessed 1 December 2009
[4]http://www.genome.jp/kegg/ last accessed 1 December 2009

It was noted that human mouse and pig genomes in KEGG was compiled from RefSeq proteins and the resulting fasta format entries were mapped to Entrez Genes, so human, mouse and pig protein fasta files were retrieved from KEGG and used for orthologue mapping of pig genes to human and mouse genes using OrthoMCL (explained later in this chapter). KEGG identifiers to Entrez Gene identifiers mapping files were retrieved from KEGG ftp service and were used in pathway mapping of gene and protein entries. For this work the KEGG release version used was KEGG release 50.0 released on 1 June 2009.

7.1.5 dbSNP

dbSNP[5] is the SNP (Single Nucleotide Polymorphism), small scale multibase deletion (IN-DEL) and short tandem repeat database by NCBI. Data submitted to dbSNP is integrated with other NCBI information sources such as Entrez gene, Entrez protein, Structure, nucleotide, taxonomy, PubMed and PubMed Central. For each of the SNP assays submitted to dbSNP, a unique identifier is created (ss number). dbSNP maps each of the submitted assays to genome and created a Reference SNP accession ID (rs number) to each of the submitted assay. All the submitted SNPs that map to the same location are clustered under one rs number. dbSNP services provide mappings from rs number to Entrez Gene identifiers.

dbSNP ftp service was used to retrieve gene allocation files for rs numbers, and the each file contains all the SNP gene allocation data for a single chromosome. These files (in XML format) contain experimental data, meta data about experiment, the nucleotide sequence accessions submitted, mRNA accession, gene accession, SNP location and 5' and 3' nucleotide sequences where the polymorphism is found, along with some flanking sequences. The files were retrieved and parsed using perl scripting to obtain rs number to Entrez Gene accession mapping. The data from *Bos taurus* dbSNP build 130 was used in this thesis.

7.1.6 OrthoMCL

OrthoMCL is used in this thesis to identify human and mouse orthologs for pig genes as satisfactory results were not obtained from databases. OrthoMCL is a graph clustering algorithm that uses sequence similarity to identify homologous proteins. It needs BLASTP and formatdb stand alone tools from NCBI. BLASTP is used for protein-protein BLAST and formatdb is used to format source data files with sequences so that it can be used by the BLAST algorithm. The algorithm works by identifying reciprocal BLAST hits (Wall et al., 2003) across two genomes by BLASTP. OrthoMCL then creates a graph, were the nodes of the graph are protein sequences and edges are pair based protein similarity scores. The edge values (edge weights) are adjusted to resolve similarity averages between the two genomes and the resulting graph is clustered using MCL algorithm to reduce

[5]http://www.ncbi.nlm.nih.gov/projects/SNP/ last accessed 1 December 2009

large clusters into smaller clusters with true orthologue relations (Li et al., 2003). Figure 7.1 illustrated the OrthoMCL workflow.

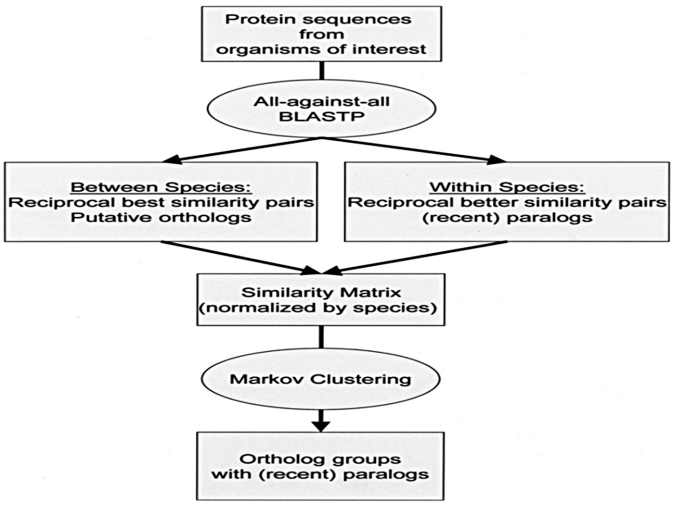


Figure 7.1: OrtoMCL workflow adopted from Li et al. (2003).

7.1.7 miRanda

MiRanda algorithm (Enright et al., 2003) was developed for the prediction of microRNAs targets in animal genome sequences. The initial use of miRanda was to predict microRNAs targets in *Caenorhabditis elegans* and *Drosophila melanogaster*. The algorithm is written in C and is open source software. It was used in this thesis to predict microRNA targets for cattle and pig genomes. The inputs needed by the algorithm are two FASTA sequence files, the first file containing the microRNA sequence and second one containing sequence of 3' untranslated region (UTR) of the gene. MiRanda works on a three phase process (see Figure 7.2), where the first one is sequence matching to assess the probability of the microRNA sequence and 3' UTR region of the gene to bind. In the matching phase the algorithm considers the effect of G-U wobble pairs (Varani and McClain, 2000), moderate insertions and deletions and rewards sequence complementarity at the 5' end of miRNA. The sequence matching algorithm is roughly similar to Smith-Waterman algorithm, but instead of accounting for sequence similarity, the algorithm accounts for sequence complementarity. Finally, four empirical rules were applied for complementary sequence matching, and the positions were counted starting from the 5' end of miRNA. The rules are (i) no mismatches allowed from positions 2 to 4; (ii) fewer than 5 mismatches were allowed between positions 3 to 12; (iii) at least one mismatch is allowed between positions 9 and L-5 , L being the total alignment length and (iv) fewer than two mismatches allowed between the last five positions of the alignment. The second phase of the algorithm is free energy calculation of optimal strand strand interaction between miRNA and UTR region. The free energy calculation was done using Vienna package (Wuchty et al., 1999), the folding routines from the Vienna 1.3 RNA secondary structure programming library

was used to determine the thermodynamic properties of a predicted duplex. The third phase of the algorithm is the determining the conservation of miRNA and target across different species. This step was done in the initial development of the algorithm to find out conserved miRNA and target pairs across different Drosophila species. In this thesis, however this step was omitted. The output file given by the algorithm contains sequence information of miRNA target and gene along with a textual representation of the match result and calculated score and free energy for the binding and the length of both the sequences. The mentioned line in the output file with score and free binding energy had the format: miRNA accession/name gene accession/name Maximum score tab Maximum free energy Total score Total free energy miRNA length 3' UTR length.

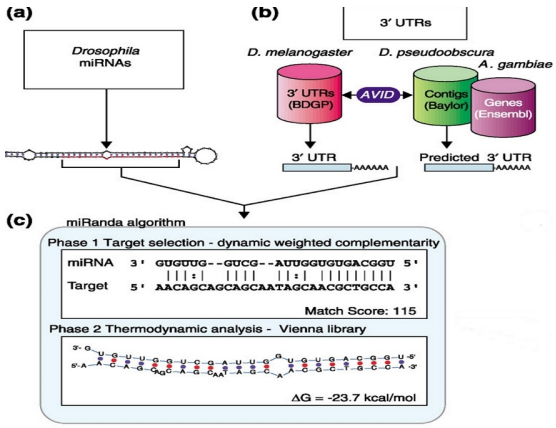


Figure 7.2: miRanda workflow adopted from Enright et al. (2003)

7.1.8 Lucene

Apache Lucene[6] is a text search engine library originally written entirely in the java programming language. Lucene is as open source project supported by the Apache software foundation[7] and released under Apache software license[8]. Indexing process is defined as the process in which a text is indexed and converted into a special format which allows rapid data search eliminating low and sequential scanning process and the output is known as index. Although the Lucene was developed initially in Java (Hatcher and Gospodneti, 2004), it was later ported into many other languages including C, C++, Perl and Python . The versatility and adaptability of Lucene made it a widely used information retrieval library. SCAIView index is Lucene based, but it supports additional functionalities such as semantic search. Additional semantic search functionalities were added to indexing (creating indexes) Lucene uses five core indexing classes which are used for creating indexes. They are:

[6]http://lucene.apache.org last accessed 02 October 2009
[7]http://www.apache.org last accessed 02 October 2009
[8]http://www.apache.org/licenses last accessed 2 November 2009

- **Index writer**
  Index writer is the central component of indexing, which creates new indexes and adds document to index

- **Directory**
  The directory class represents the location of the index. The class is abstract, which allows its subclasses to store index where they fit.

- **Analyzer**
  Analyzer extracts tokens to be indexed from the text and eliminates others. This class also has implementations to skip the stop words and case sensitivity.

- **Document**
  Document class is defined as a collection of fields. This class can be considered as a collection of data that need to be retrieved at a later time.

- **Field**
  Field is a piece of data that is either queried against or retrieved from the index during a search.

Lucene uses inverted index concept (see Figure 7.3), where the tokens extracted during the index creation is used as look up keys and documents are not considered as central entities. In the index, documents with different sets of field can co exist. Lucene also provides the concept of appendable fields, where synonyms of a word can be appended together into a single field and is then used to create a Lucene field. This property of Lucene makes it highly adaptable for the use indexing documents containing gene and protein data. For retrieving the documents associated with a search term, Lucene uses its search classes. The Lucene search classes are:

- **Index searcher**
  Index searcher is the central component of index search, a method that exposes several search methods and opens an index for reading puprpses.

- **Term**
  Term is the basic unit of searching, it consist of a pair of string elements, name of the field and value of the field.

- **Query**
  Query is an abstract parent classes for different query types.

- **Term query**
  Term query is the most basic type of query used by Lucene, and is a primitive query type. Term query is used to match documents that contain specific values.

- **Hits**
  Hit class is used to rank search results and is a container of pointers.

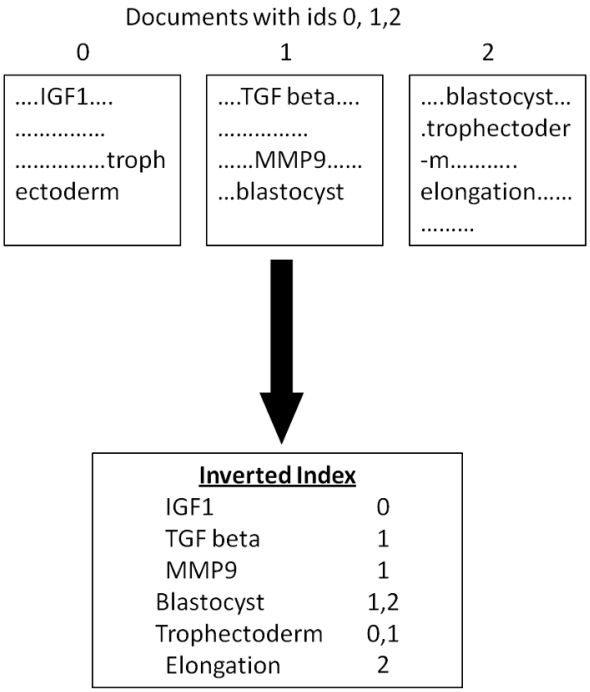


Figure 7.3: Lucene inverted index example

7.1.9 Protégé and Knowtator

Protégé[9] (Gennari et al., 2003) is an open source ontology editor and a knowledge base framework. Knowtator[10] (Ogren, 2006) is a general purpose text annotation tool integrated to the Protégé representation system as a plugin. Both Protégé and the text annotation tool Knowtator was used in this thesis for the annotation of test corpus for the dictionaries. The test corpus was derived through keyword search experiments in PubMed. Protégé is a java based extensible environment and provides support for plugins. Protégé was initially designed as an application to develop knowledge acquisition tools for medical planning (Noy et al., 2001). A Protégé knowledge base comprises of classes, instances, slots and facet frames. Classes represent the concepts of a domain and are organized hierarchically. Instances represent individual entries in a class and slots define the relationships, with a class and instances or among classes. Facets represent the values that each slot can hold. Knowtator is designed to use the knowledge representation capabilities of Protégé for annotation of complex schema. One key feature of Knowtator is its pluggable infrastructure that can handle multiple data formats of texts and the original text used is not modified during the annotation procedure. For corpus annotation, Knowtator creates three sets of annotation files (.pins file, .pont file and .pprj file) in which .pins file contains the annotated entities along with spans and the other two files

[9]http://protege.stanford.edu last accessed 05 October 2009
[10]http://knowtator.sourceforge.net last accessed 05October 2009

include annotation specification and other information for the user. The .pins file was parsed using a perl script to extract the annotation.

7.1.10 Cytoscape

Cytoscape[11] (Shannon et al., 2003) is an open source bioinformatics platform that is used to visualize molecular interaction networks. The tool was uses in this thesis to visualize various interaction networks retrieved from SCAIView and validate the predicted networks using experimentally predicted interaction data. Gene expression profiles and other confirmed experimental data can be integrated into Cytoscape. It is most commonly used with protein-protein, proteingene and gene-gene interactions. Cytoscape provides basic functionalities for network visualisation, integration of various experimental data to the networks and linking network nodes to external database.

7.1.11 ProMiner

ProMiner (Fluck et al., 2003) is a named entity recognition system that is already being used the human, murine and Arabidopsis genomics domain. The system can be divided into three major parts with the generation and curation of dictionaries being the first. In this step, names and synonyms of biological entities, from different database resources are assembled into a dictionary. Each name and synonym is considered as a token, or association of more than one token. Tokens are defined as small bits of strings, for example, 'matrix metalloproteinase 9' can be split into three tokens 'matrix', 'metalloproteinase' and '9' and the problem of biological entity recognition is solved at the token level. The set of all the identified tokens are separated into token classes. The token classes are based on varying degrees of significance that tokens have during occurrence detection. The curation of the dictionaries was also done at this stage. The main idea behind dictionary creation is the strict removal of synonyms that are unspecific and creation of a comprehensive dictionary. This stage also involves the preprocessing of the synonyms. Acronyms and subtype specifiers (example: a- indicating alpha) are expanded and insertion of space is permitted between words and digits (example MMP9 and MMP 9). Certain rules are also employed in this stage, which would account for the case sensitivity and detection of previously unidentified synonyms.

Dictionary generation step involves rule based classification of synonyms as well. As the first step in this process, Porter stemmer was used to count occurrences of all stemmed words in all abstracts in the MEDLINE database. Based on certain criterion all the words were combined into three different groups. The most frequently occurring synonyms are classified as "unspecific synonyms". This classification is based on the principle that "frequently occurring terms are less likely to be used as protein names". The synonyms which are identified through regular expression in the previous step are also grouped in this class. The synonyms in this class are used for the disambiguation of other matches. The synonyms in the class "unspecified synonyms" are used as synonyms when certain

[11] http://www.cytoscape.org last accessed 23 September 2009

context specific words like protein, gene and transcripts are used along with them. The second class is the "case-sensitive synonyms" where the entries in the class could be differentiated from another dictionary entry only if the case of the synonym is considered. The third class of synonyms constituted all the other entities which could be identified during an approximate case insensitive manner.

The second step in the process is the search procedure. The biological entities in the texts are identified by a text wide search and processing one token at a time and a set of possible solutions (candidate solutions) are considered for each biological entity. The search algorithm employs two different scoring schemes; a boundary score and an acceptance score for matching a biological entity with a candidate solution. Boundary score calculates the number of mismatched tokens in a candidate solution and acceptance score is a linear combination of match and mismatch terms that are token specific. Match and mismatch terms are described as the percentage of matched tokens in a given class and the number of additionally found tokens during a candidate extension. When appropriate weighting is given, the acceptance score is powerful enough to identify the different variations of a synonym and ignore false substring matches. Different search methods are employed for synonyms in different synonym class. An exact search method allowing no deletion, insertion or permutation is applied for synonyms of the "unspecific class". In case of synonyms consisting of more than two tokens, an approximate search method that permits deletion, permutation and insertion of tokens is employed. The synonym class determines if the search has to be done in a case sensitive manner or case insensitive manner. This step includes identification of previously undefined abbreviations as well.

Match filtering is the final step, in which unspecific synonyms and overlapping matches are filtered out. During the search process, if only one synonym is found as match for certain text position, the synonym is accepted as a hit. But, it can also happen that a set of synonyms are found as a hit for a certain position in the text. In such cases the synonym match with a higher acceptance score is considered as a match. This matching procedure would also accept the synonyms of the class "unspecific synonyms" as a match, it the synonyms are accompanied by terms having a high acceptance score. In case of synonyms which can be mapped to more than one gene name, the synonym is mapped to the gene/protein entry for which additional synonyms are found in the same abstract. The match filtering stage includes an organism filtering stage as well, which would filter the abstracts based on organism name (Hanisch et al., 2005). A complete workflow of entity recognition process by ProMiner is given in Figure 7.4

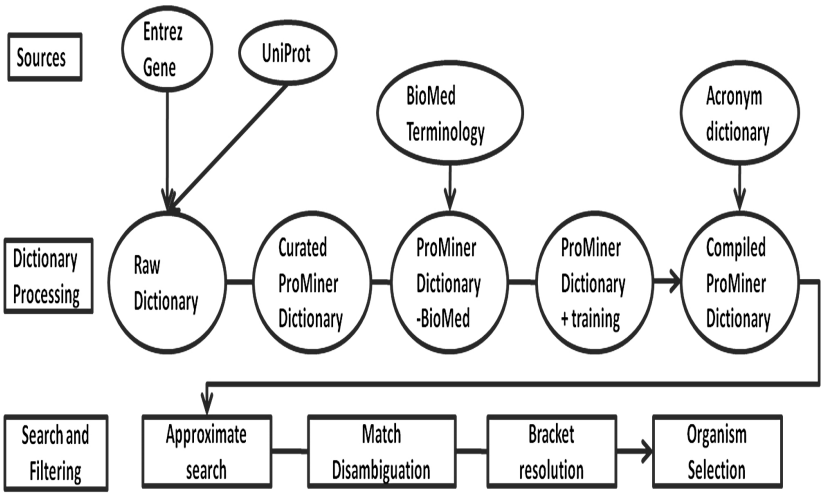


Figure 7.4: ProMiner workflow

ProMiner participated in the BioCreAtIvE challenges and the results of the BioCreAtIvE challenges are give in Table 7.1. ProMiner also participated in the BioCreAtIvE II challenge, where the task was to link Entrez Gene identifiers to gene mentions in the abstract. The system was adopted for the challenge using human gene dictionaries and incorporating an initial step in which spelling variants were automatically added to the dictionaries and external dictionaries were used for the removal of non-gene names. The ambiguous gene mentions in the abstracts were identified by using the dictionary, an abbreviation dictionary or depending on the frequency occurrence of words in the text Morgan et al. (2008). The detection of ambiguous gene names was done inorder to improve the precision of the system. The dictionary generated through automatic addition of spelling variants were filtered through regular expressions or using a biomedical terminology dictionary created based on different OBO ontologies for disease, tissue, organisms and protein family names, generated manually. Additionally, an acronym dictionary containing gene specific short forms and non gene specific long forms were used for dictionary disambiguation in this compilation step. The results of the BioCreAtIvE II is given in Figure 7.5

Table 7.1: ProMiner performance adopted from Karopka et al. (2006)

Organism	Precision	Recall	F-score
Mouse	0.77	0.81	0.79
Yeast	0.97	0.84	0.90
Fly	0.83	0.80	0.82
Fly Accept matches of synonyms associated to up to 3 different Entrez Gene entries	0.74	0.83	0.79
Human	0.86	0.81	0.84

	Test Run 1 D1, O-	Test Run 2 D3, O-	Test Run 3 D1, O+	Test DictOrig D1, O-	Test DictSub D1, O-	Train D1, O-	Train-brackets D1, O-	Test-brackets D1, O-
F-measure	**0.799**	**0.790**	**0.779**	0.792	0.847	0.784	0.776	0.799
Recall	**0.768**	**0.803**	**0.730**	0.777	0.811	0.755	0.736	0.768
Precision	**0.833**	**0.779**	**0.835**	0.809	0.885	0.819	0.820	0.833
Quartile	**1**	**1**	**1**					

Figure 7.5: ProMiner results of BioCreAtIvE II challenge adopted from Fluck et al. (2007). Legend: Test Run -Test corpus run; D1, D3 - Disambiguation thresholds; O+, O- Organism selection; DictOrig - BioCreAtIvE training corpus, DictSub - genes gold standard.

7.1.12 SCAIView

"Building on existing named entity recognition technology and ontologies, SCAIView exploits literature mining to enable both hypothesis generation and biological discovery" (Gattermayer, 2007). SCAIView (see Figure 7.6) is an advanced semantic based search engine which was materialized as a part of the European Union funded @neurIST[12] project. SCAIView is developed as an integrated knowledge environment system (Hofmann-Apitius et al., 2008) which is designed for the identification of the genes that correlate with the disease aneurysm (characterized by balloon like bulging of the blood vessel) by the semantic analysis of PubMed abstracts.

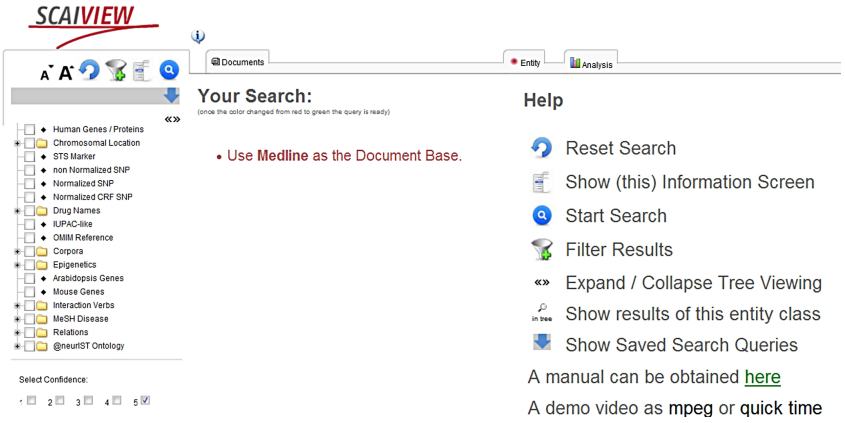


Figure 7.6: SCAIView homepage

The @neurIST project has four main end user application suites namely @neuLink, @neuFuse, @neuRisk and @neuEndo. SCAIView was initially developed as a part of @neuLink, an application suite that links genetics to disease via Knowledge Discovery and the data mining tasks that can be accomplished through @neuLink include (Friedrich et al., 2008a) :

[12]http://www.aneurist.org last accessed 13 September 2009

- Finding candidate genes and gene variants that are related to the disease (this part of @neuLink is known as SCAIView),
- Finding disease associated proteins through visualization and analysis of protein-protein interaction networks.
- Analysis of gene expression data and result integration.
- Data mining the @neuIST database for generic risk factors.

Text mining methods are used to find candidate genes and their variants in text data. Taking the necessary text mining time into consideration, the process is distributed over the grid service used by the @neurIST project, several computer clusters and project partners. The unstructured text mainly used is PubMed abstracts and the system allows semantic search for PubMed abstracts. The results are stored in a layer, which forms the base for "Find Candidate gens and variations" service of the @neuLink service. Besides text mining data, the layer also used data from SRS[13], dbSNP, Entrez Gene, UniProt and DrugBank[14]. The final result presented to the end user after the integration of all the results (see Figure 7.7).

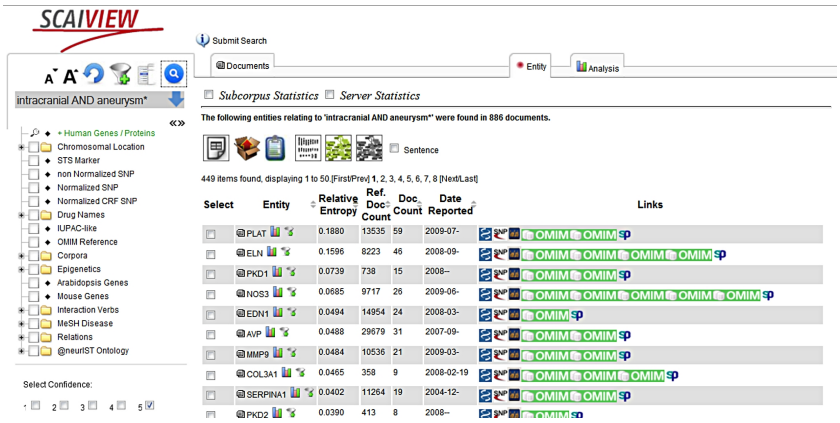


Figure 7.7: SCAIView entity view

The second step involves visualization and analysis of candidate proteins and interaction partners that can be responsible for the disease. The candidate genes are obtained from the "Find Candidate gens and variations" and possible interaction partners and interactions are retrieved from the PIANA[15] (Aragues et al., 2006) database, which is a collection of

[13]http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession last accessed 13 September 2009
[14]http://www.drugbank.ca last accessed 13 September 2009
[15]http://sbi.imim.es/piana/ last accessed 13 September 2009, the PIANA system has been updated to BIANA.

data from some of the publicly available databases like IntAct[16], MINT[17], and BIND[18]. The interaction networks obtained can be accessed through data access service and analyzed using network analysis tool Cytoscape (Shannon et al., 2003). In @neuLink, Microarray gene expression data is used to confirm the knowledge discovery results that are obtained through text mining. A service oriented microarray workflow is used to integrate microarray results to @neuLink, the workflow is based on Bioconductor tools and can access data from Array Express[19] and Gene Expression Omnibus[20] (GEO) along with other Minimum Information About Microarray Experiments (MIAME) compliant datasets.

The data mining module of @neuLink is used to mine the @neurIST patient database. The module uses standard tool provided by Bioconductor and R. Besides mining the database the module is also aimed at finding association between the disease and non genetic factors that are available in the @neurIST database. Figure 7.8 shows SCAIView workflow.

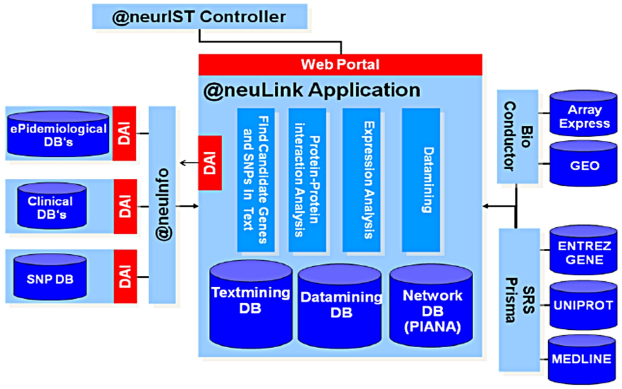


Figure 7.8: SCAIView work flow adopted from Friedrich et al. (2008a)

An additional feature of SCAIView is its aneurysm ontology and ontology based searching. In the @neuLink interface, the ontology has three main functions to serve: As a terminology for the project, Supporting semantic analysis through relations between entities Support data integration. The ontology merges different scientific disciplines and views on aneurysms and subarachnoid hemorrhage. The relevant term and concepts in the aneurysm ontology are compiled from clinical databases, literature, knowledge from domain experts, UMLS (Unified Medical Language system) Meta thesaurus and public domain databases for molecular biology and adapts several existing ontologies available. The aneurysm ontology is modeled as a "functional ontology", which can enable

[16]http://www.ebi.ac.uk/intact/main.xhtml last accessed 13 September 2009
[17]http://mint.bio.uniroma2.it/mint/Welcome.do last accessed 13 September 2009
[18]http://bond.unleashedinformatics.com last accessed 13 September 2009, BIND is now BOND (Biomolecular Object Network Database)
[19]http://www.ebi.ac.uk/microarray-as/ae last accessed 13 September 2009
[20]http://www.ncbi.nlm.nih.gov/geo last accessed 13 September 2009

communication and corporation between different layers in @neuLink. The aneurysm ontology is implemented in a dictionary based approach that describes the relevant terms describing the disease intercranial aneurysm and associated concepts. ProMiner was used as to identify all the terms in ontology in the given corpus. An advantage of converting the ontology into a dictionary format is that for ontology- terms that are identified in text, the database reference to the term could be given (UMLS, GO etc). In SCAIView the ontology is displayed in an XML (eXtended Markup Language) tree based format with hierarchical classification of the terms. SCAIView provides a relative ranking of the entities based on relative entropy score (Friedrich et al., 2008b). The ranking system based on relative entropy score,(Kullback/ Leibler divergence) (Kullback and Leibler, 1951), where the entire reference corpora used in SCAIView is compared against a subset corpora obtained through a search in SCAIView and the entities are ranked accordingly. Further addition to SCAVIew is a full text version of SCAIView, the full text version has the advantage on the information content.

7.2 Methodology

The methodology section describes the main methods used in thesis for data retrieval, generation and manipulation. The subsections in this section are:

- Generation and curation of species specific gene and protein name and microRNA dictionaries.
- Mapping of external database information to ProMiner results.
- MicroRNA target analysis and mapping of computationally determined microRNA results to dictionaries.
- Terminology analysis of Cattle preimplantation period.
- Indexing
- Corpus annotation and performance evaluation

7.2.1 Generation and curation of species specific gene and protein name and microRNA dictionaries.

The very first step in the adaptation of SCAIView was the generation of gene and protein name dictionaries, specifically for cattle and pig. A set of dictionaries had to be created for each organism. The first one termed as the “curated dictionary“ was generated using the gene and protein data that is available for each of the organism. For the generation of these dictionaries, the primary steps are data assembly and data generation, which are explained here. The data for the generation of curated dictionary were assembled from the UniProt database. The second dictionary, termed as “generated dictionary“ was created by mapping cattle and pig genes to their orthologues in human and mouse

genome, considering them as model organisms. The reason for creating an orthologue dictionary using human and mouse gene data is the quality of gene annotation in human and mouse genome and the sequence similarity of these genome to cattle and pig genome (refer to Figure 7.9).

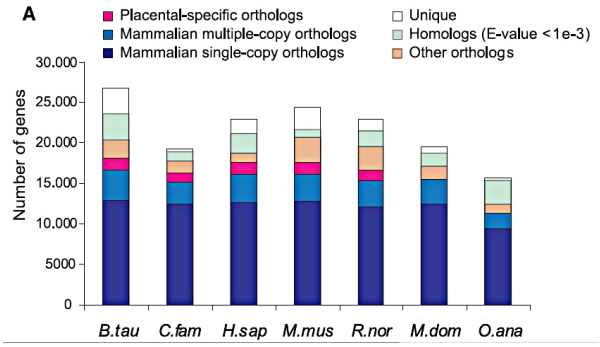


Figure 7.9: Genome similarity among organisms Elsik et al. (2009)

Own work

For generating curated dictionaries for cattle and pig, UniProt data for both of the organisms were used. Further, additional spelling variants and synonyms for each of the entries was added during a later stage, this data also included data from the Entrez Gene for all those UniProt entries that could be mapped to Entrez Gene entries. The cattle dictionary has approximately 13,800 entries and pig dictionary has 7,100 entries. Different strategies were used for the orthologue mapping of cattle and pig genome. Since cattle genome sequencing is complete and a well annotated list of cattle genes is available from public databases, orthologue mapping service of the Ensembl database was used for this purpose. Ensembl integrates the Biomart[21] system for data querying. The system allows large scale querying and allows the user to export the data into different file formats. For retrieving the orthologue mapping, the latest version of Ensembl database at that time was selected from the pull down menu, and the selection resulted in another pull down menu with the genome builds of all the organisms that were present in the database. From this list, *Bos taurus* genome was selected and the list was limited to all the genes with corresponding Entrez Gene identifiers. This page also includes an option for multispecies comparisons. From the list, orthologue human genes were selected. The next step was the selection of attributes that needed to be retrieved. For cattle and human genes, Ensembl gene id was selected as the attribute as orthologue retrieval in Ensembl Biomart allows the genes only to be mapped to Ensembl identifiers and not to any other database identifiers. In this way a tab separated table with cattle Ensembl gene identifiers and human Ensembl gene identifiers was retrieved. The same process was repeated for cattle mouse orthologs and the final result was obtained by merging the cattle-human orthologues and cattle-mouse orthologues. For all the genes with cattle

[21]http://www.biomart.org last accessed 13 September 2009

and mouse orthologues cattle-human orthologue data and cattle-mouse orthologue data were merged to cattle-human-mouse orthologue data. Orthologue data for approximately 17,500 cattle genes were retrieved in this way. Since pig genome data was not available in Ensembl database or in most of the major public domain databases at that time, other orthologue specific databases were searched for human and murine orthologues for pig. The searched databases include EGO[22], OrthoMCL[23] (Roos et al., 2006), InParanoid[24], orthologue mapping in MGI database[25] (Blake et al., 2003b). Only MGI database provided orthologue mapping of pig genes to human and mouse counterparts, but the output result was too few, roughly around 600 pig genes were mapped to human and murine orthologs. So an alternate approach was used to map pig genome to human and mouse orthologues. An orthologue detection algorithm, OrthoMCL (refer to Section 7.1.6) was used to identify model organism orthologs for pig. The data supplied to the algorithm included around approximately 21,000 human genes, 9,000 mouse genes and 7,000 pig genes in the form of three different protein FASTA files and default parameters in the algorithm were used. The run produced two different files; the first one included the clustering output given by OrthoMCL and the second one included pair wise BLAST result for all the proteins in all the three files in BLAST tabular BLAST result format (m8 format (refer to Figure 7.11)). A post processing step was needed for the result from the OrthoMCL clustering results, as some of the cluster groups had a large number of genes in them (upto $\sim$50 genes in some cluster group). The post processing step was done in different steps, using perl scripts. As a first post processing step, the gene identifiers of all the pig genes were extracted from the OrthoMCL clustering result using a parser. By keeping the clustering result from OrthoMCL as a seed point, all the BLAST scores where the pig genes were used as the source sequence and human and mouse genes as the target sequences were extracted from the second OrthoMCL output file (BLAST result). All the BLAST results for individual pig genes were clustered into one, keeping the original format of the BLAST results. After the clustering, each of the clustered groups was treated as 2 dimensional matrices and matrix sorting was performed on individual matrices based on sequence identity of the source and target proteins of the BLAST result. Based on the sequence identity of the source and target sequences and all the pairs with less than 35% of the sequence similarity were filtered out, based on twilight zone of protein sequence similarity (Rost, 1999).

[22]http://compbio.dfci.harvard.edu/tgi/ego last accessed 13 September 2009
[23]http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi last accessed 13 September 2009
[24]http://inparanoid.sbc.su.se/cgi-bin/index.cgi last accessed 13 September 2009
[25]http://www.informatics.jax.org last accessed 13 September 2009

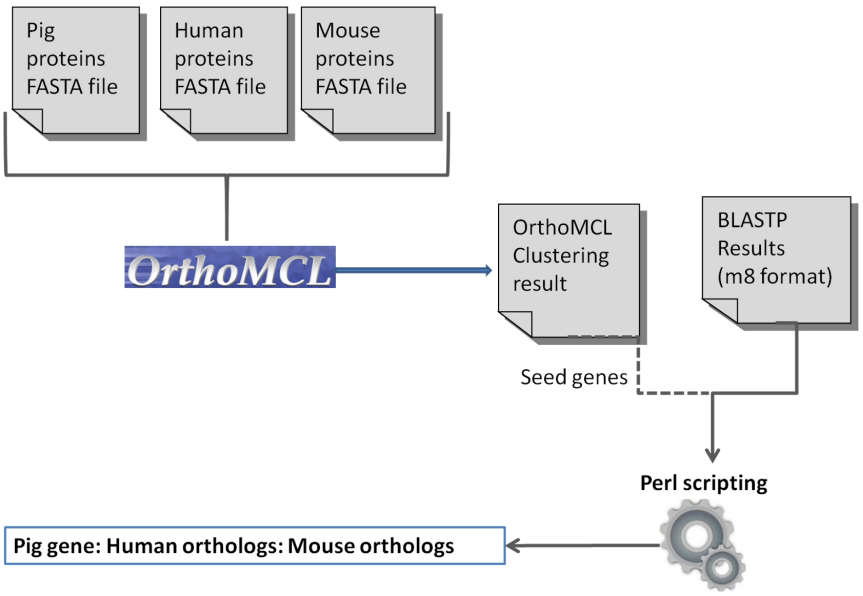


Figure 7.10: Workflow used for identifying human and mouse orthologs for pigs

Query id	Subject id	% identity	alignment length	mismatches	gap openings	q. start	q. end	s. start	s. end	e-value	bit score
hsa:390053	ssc:100155934	32.5	224	148	2	7	230	5	225	4.00E-24	109

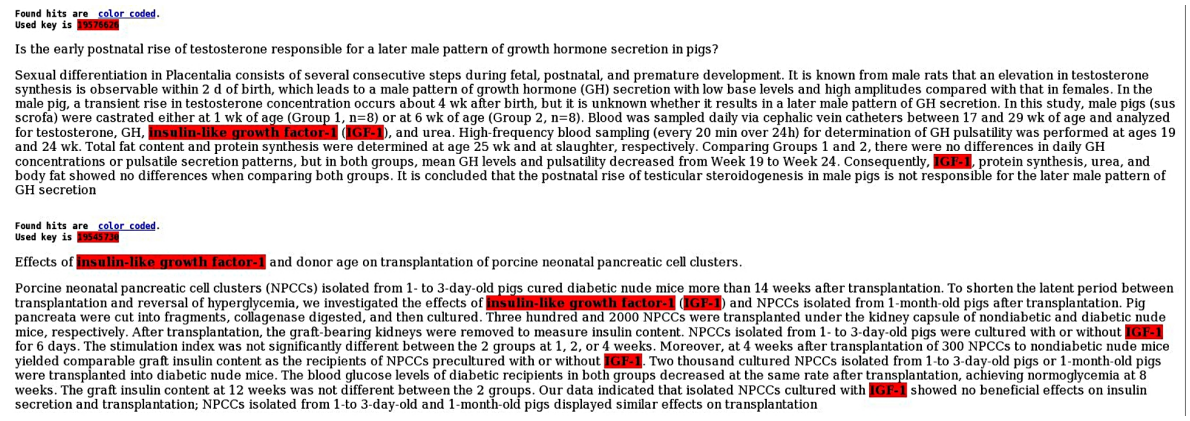
Figure 7.11: m8 format description using an example

The later steps in the processing steps include further filtering of the obtained results. After the filtering of results based on the sequence similarity, the matrix was further sorted based on the bit score of the BLAST result. The top scoring bit score after the scoring was selected as a standard and all the results with BLAST scores and sequence identities that fell within a certain limit of the standard BLAST score and sequence identity were considered as orthologs for the particular pig gene. The resulting human and mouse genes were mapped to the source pig gene, and data from Entrez Gene and Uniprot database were used to create the orthologue pig dictionary.

Curation of Dictionaries

The second step in creating protein and gene name dictionaries was the formatting of the assembled/generated data to the ProMiner dictionary format and submitting those dictionaries for a ProMiner run on abstracts (complete MEDLINE abstracts or a subset specified by the user). Here, the dictionary curation steps done for cattle and pig dictionaries are described. For using the generated dictionaries in ProMiner for named entity recognition, the entries in dictionaries need to be in a special format. The dictionaries were converted into ProMiner format by using perl scripting (ProMiner

dictionary format example: 768107@ENTREZGENE|Regucalcin:Regucalcin|Senescence marker protein 30|Senescence marker protein30|SMP-30|SMP30|SMP 30). The idea behind the format is normalization, where all the gene names representing the same gene are mapped to a unique identifier (here to a unique database identifier from Entrez Gene or UniProt). Along with mapping to external database identifier, ProMiner creates its own internal identifiers so that a gene/protein entry and associated synonyms are mapped not only to a unique database identifier but also to a unique internal identifier. The created internal identifiers are used for mapping purposes at a later stage in the process. The format- converted gene and protein name dictionaries are subjected to a run on ProMiner on approximately 470,000 MEDLINE abstracts (the total number of abstracts in MEDLINE relating to cattle and pig). ProMiner accepts text in two kinds of input formats, either as free text in files or as list of PMID identifiers for MEDLINE abstracts (as .ids file). When a list of PMID identifiers are given, ProMiner retrieves the abstracts using PMID from its internal database, and in this thesis since the results needed to be indexed using Lucene for SCAIView integration, a SCAI custom built PubMed abstract format, lucmed was used. In lucmed format the heading sections of the abstracts are avoided and PMID is used as the unique identifier, added to the main body of the abstract. For generation and curation purposes of cattle and pig gene/protein name dictionaries, a list of the above mentioned MEDLINE abstracts were given. ProMiner provides a visualization interface for analyzing the run results. The interface shows the found entities, synonyms of the found entities, total number of occurrences of the entity and the abstracts in which the entity was found. The visualization interface was used to manually identify false positives in the later curation stages (see Figure 7.12).


Found hits are color coded.
Used key is 19575824

Is the early postnatal rise of testosterone responsible for a later male pattern of growth hormone secretion in pigs?

Sexual differentiation in Placentalia consists of several consecutive steps during fetal, postnatal, and premature development. It is known from male rats that an elevation in testosterone synthesis is observable within 2 d of birth, which leads to a male pattern of growth hormone (GH) secretion with low base levels and high amplitudes compared with that in females. In the male pig, a transient rise in testosterone concentration occurs about 4 wk after birth, but it is unknown whether it results in a later male pattern of GH secretion. In this study, male pigs (sus scrofa) were castrated either at 1 wk of age (Group 1, n=8) or at 6 wk of age (Group 2, n=8). Blood was sampled daily via cephalic vein catheters between 17 and 29 wk of age and analyzed for testosterone, GH, insulin-like growth factor-1 (IGF-1), and urea. High-frequency blood sampling (every 20 min over 24h) for determination of GH pulsatility was performed at ages 19 and 24 wk. Total fat content and protein synthesis were determined at age 25 wk and at slaughter, respectively. Comparing Groups 1 and 2, there were no differences in daily GH concentrations or pulsatile secretion patterns, but in both groups, mean GH levels and pulsatility decreased from Week 19 to Week 24. Consequently, IGF-1, protein synthesis, urea, and body fat showed no differences when comparing both groups. It is concluded that the postnatal rise of testicular steroidogenesis in male pigs is not responsible for the later male pattern of GH secretion

Found hits are color coded.
Used key is 19545736

Effects of insulin-like growth factor-1 and donor age on transplantation of porcine neonatal pancreatic cell clusters.

Porcine neonatal pancreatic cell clusters (NPCCs) isolated from 1- to 3-day-old pigs cured diabetic nude mice more than 14 weeks after transplantation. To shorten the latent period between transplantation and reversal of hyperglycemia, we investigated the effects of insulin-like growth factor-1 (IGF-1) and NPCCs isolated from 1-month-old pigs after transplantation. Pig pancreata were cut into fragments, collagenase digested, and then cultured. Three hundred and 2000 NPCCs were transplanted under the kidney capsule of nondiabetic and diabetic nude mice, respectively. After transplantation, the graft-bearing kidneys were removed to measure insulin content. NPCCs isolated from 1- to 3-day-old pigs were cultured with or without IGF-1 for 6 days. The stimulation index was not significantly different between the 2 groups at 1, 2, or 4 weeks. Moreover, at 4 weeks after transplantation of 300 NPCCs to nondiabetic nude mice yielded comparable graft insulin content as the recipients of NPCCs precultured with or without IGF-1. Two thousand cultured NPCCs isolated from 1-to 3-day-old pigs or 1-month-old pigs were transplanted into diabetic nude mice. The blood glucose levels of diabetic recipients in both groups decreased at the same rate after transplantation, achieving normoglycemia at 8 weeks. The graft insulin content at 12 weeks was not different between the 2 groups. Our data indicated that isolated NPCCs cultured with IGF-1 showed no beneficial effects on insulin secretion and transplantation; NPCCs isolated from 1-to 3-day-old and 1-month-old pigs displayed similar effects on transplantation

Figure 7.12: ProMiner visualization interface used for curation

The initial run was done as a test run to identify all the possible false positives that could be included in the raw dictionary. Some examples of the false positives are: gene/protein names that are used in context of more than one gene name, acronyms of gene names, which have more than one expansion, certain organism names that are included as a part of the gene name, or included to denote the organism in which the gene was identified at first and some commonly occurring words that are used as gene/protein

name. An example is the word “CAT“, which in some abstracts, is used to denote the animal cat, in some other abstracts used for the gene CAT (Entrez Gene ID: 531682) and in other abstracts used for the bacterial enzyme Chloramphenicol acetyltransferase (CAT). Similar is the case with acronyms, the acronym IVD is used as a gene name IVD (isovaleryl Conezyme A dehdrogenase, Entrez Gene ID: 510440), also as an acronym for in vitro derived (IVD) as well as for intervertebral disc (IVD). But in this problem of acronym disambiguation can be solved thorough Latent Semantic Indexing (LSI) (Oster, 2008). Some of the other noises in the gene/protein name dictionary include cDNA information of the transcript from which the gene is annotated and some gene name variants for isoforms which rarely occur in the text. These false positives and noises had to be filtered from the dictionary. ProMiner allows the user to remove the synonyms or false positive terms based on the context. For instance, ambiguous names which occur as synonyms for a number of gene/protein name. Some words create ambiguity only within the limit of their own identifiers; these words were removed on that specific context. Take the case of acronyms with more than one expansion, in case of these acronyms, the ambiguity is only within the limit of a range of gene/protein name for the acronym. In some conditions, the entire synonyms and the identifier were removed as a result of redundant occurrences. These redundant occurrences were especially noticeable in case of TrEMBL data. In certain instances it was noted that certain spelling variants of the gene/protein name and some synonyms were only found in texts and not in databases. To illustrate, for the gene insulin-like growth factor 1 the normal acronyms found in databases are IGF-1 and IGF-I, but in texts it can be seen that “IGF1“ is also used as a synonym for the gene (PMID 18586434). ProMiner has an allocated file in which the entries are either the false positives that needs to be removed or the synonyms that need to be added to an identifier. The identifier, followed by ‘:‘ and ‘-‘ or ‘+‘ is used in front of the term to be removed or added depending on the situation. There exists certain case sensitive gene and protein names, which can be categorized as gene/protein names when the synonym is represented entirely in capital letters or a mix of capital and small letters (STAR and StAR :- steroidogenic acute regulatory protein Entrez GeneID: 782522) and the meaning of these terms completely changes when used only in small letter. These words were treated in a special manner in the dictionary curation. ProMiner has two dedicated files which would consider the entries without word permutations, insertion of spaces or change of cases. Curation was done as an iterative process and was the most important step in this section.

ProMiner entity visualization was used to visualize the entites identified by cattle and pig dictionaries. The falsely identified entities (some words that are not gene/protien mentions, but marked by the system as a gene/protien entity) and entities that were not identified (false negatives) were manuall identified and were used as inputs for the compilation and curation run. All the ambiguous synonyms and acronyms that were to be removed from the system along with the synonyms that were to be added to respective identifiers were given as inputs for in the file. The generated gene/protein name dictionary was then compiled with ProMiner using the mentioned files as inputs. The compilation step accounted for the removal of ambiguous terms and addition of synonyms supplied to the system. The dictionary compilation step also created two other

files namely the .map file, which was later used as a starting point for external database mapping and a .single file which was gave a list of all the single word synonyms where no permutation of the synonyms were allowed as specified by the user. In this project, the entire ProMiner dictionary curation and compliation steps were used. A generalized workflow of the curation step is given in Figure 7.13.

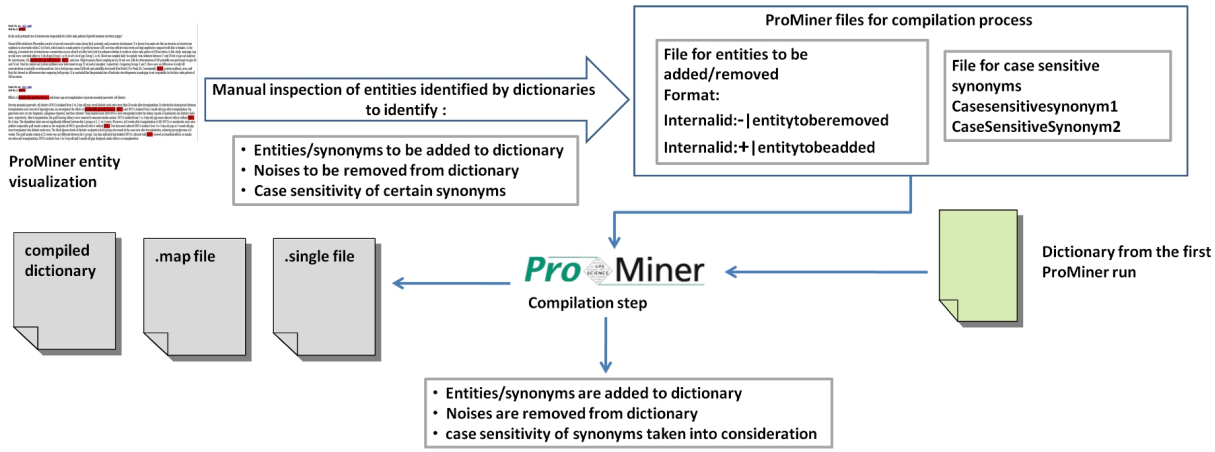


Figure 7.13: Curation process workflow

The compiled dictionary was then given as the input gene/protein name dictionary for a ProMiner run on the same 470,000 abstracts. The output of a ProMiner run is a .prt file which stores the metadata on the searched abstracts. A PRT file is a tab separated file with the columns representing the information of the entities extracted from the abstracts (see Table 7.2). A prt file contains a total of 37 tabs. The first column in the abstract gives the PMID of the abstracts which were used for the search run, followed by the internal identifier of the found gene/protein name, the next three columns gives the identified term in three different formats: the normalized term; the term to which the synonym is mapped and the spelling variant identified in the abstract, spans of the term (word count from beginning of the term to end of the term in a document) in the abstracts and other details. The .prt files obtained were indexed with SCAIView indexing process, specifying the PMIDs of the abstracts in a separate .ids file for restricting the indexing process to those abstracts thus making the indexing specific for cattle and pig genes. The indexing process was a major step in the adaptation of SCAIView, since the specified index is the major gene protein name data source for SCAIView.

Table 7.2: Some of the important columns in a PRT file, using examples from cattle dictionary

Columns:	column 1	column 2	column 3	column 7	column 8	column 10
Designations:	Document id (PMID)	ProMiner Internal identifier for entity	normalized entity	entity occurrence in document	span beginning	span end
Examples:	19634707	BTC_005874	tnfalpha	TNFalpha	16	24
	19524660	BTC_005259	eNOS	eNOS	1086	1090
	19633431	BTC_005415	cdc2	Cdc2	196	200
	19633132	BTC_003644	gfra1	GFRA1	874	879

MicroRNA dictionaries

For generating microRNA dictionaries microRNA data file in EMBL format[26] from microRNA database (miRBase) was retrieved and parsed using a perl script written for the purpose. The file contained microRNAs data for all the organisms with microRNAs sequences in the miRBase database. But for the purpose then microRNA data for cattle, human, mouse and pig were parsed. The data in the retrieved file followed microRNAs naming conventions. But it was found that in certain text documents authors often introduce spelling variants for microRNAs names. For example, for the human microRNA miR-21 (hsa-miR-21) terms like MicroRNA-21 (found in PMID 196892430), microRNA 21(found in PMID 19547998) miRNA-21 (found in PMID 19450585) were found in abstracts. So appropriate changes were made in the perl parser script to include all the mentioned terms while generating the dictionary. MicroRNA dictionaries were created for cattle, human, mouse and pig. The created microRNAs dictionaries were used as base dictionaries for ProMiner run on the entire Veterinary corpus. Since the nomenclature and mentioning of microRNAs are unambiguous no curation step was done for microRNA dictionaries. Finally, all the PRT files created from ProMiner run on were used as data sources for the indexing process with SCAIView.

7.2.2 Corpus annotation and performance evaluation

For performance evaluation of the two sets of dictionaries, two sets of corpus were annotated using Knowtator plugin in Protégé. The annotated set of corpora was considered as the ‘golden standard‘ to which the performance of the dictionaries were compared.For result analysis two sets of corpus were generated, Frst corpus covering a wide range of abstracts from cattle and pig genomics and second corpus, covering a specifc range of cattle and pig abstracts concerning preimplantation and meat quality.

[26]http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html last accessed 2 November 2009

General corpus generation

For the creating a general test corpus, cattle and pig PMIDs were retrieved from PubMed using keyword search. For cattle corpus generation, the query "cattle AND genome" retrieved a list of 10347 PMIDs. From this list, 200 PMIDs for the test corpus were retrieved. The first 100 PMIDs were at random, in sets of 10-15 PMIDs from the first 5000 PMIDs and the second 100 PMIDs were retrieved from the last 5000 PMIDs in the same way. To generate a test corpus for pig dictionaries, the query term used was "pig AND genome" and the query retrieved a list of 7147 PMIDs. From the list 200 PMIDs were retrived in th same way a cattle corpus PMID list. Using these PMID lists, abstracts were retrived from the internal database in lucmed format.

Specifc corpus

For specifc corpus, a list of pmids relating to cattle preimplantaion and pig meat quality were generated using PubMed keyword search. The search strategies were specifc (for example"cattle AND preimplantation" and "pig AND meat quality OR carcass quality genetics") The initial PMID lists for the two corpora prepared for annotation as a part of this thesis contained about ∼400 PMIDs each and 200 PMIDs were selected from each of the corpora at random. Since the corpus was biased towards these fields, the documents contained more number of gene protein entities when compared to the general corpus.

The gene and protein entities found in all the abstracts were annotated using Protégé and Knowtator. For annotation, the guideline selected was that "no hormone names and no gene family name were to be annotated". The annotated gene and protein entities in each of the abstracts along with the corresponding PMIDs and the word span (word count starting from the beginning of the gene/protein entry to the end of the same entry) were retrieved from .pins file in the Knowtator annotation directory using perl scripting. The same extracted PMID lists were given as the base text source for ProMiner test run. For all the cattle dictionaries the PMID list of 200 abstracts were used and for pig dictionaries, 200 pig PMID list was used. The gene and protein entry, word span and corresponding PMID were retrieved from the corresponding PRT files and gene and protein entities along with the word spans identified in each abstract by ProMiner was compared with gene protein entities and the word span of each entry in each of the abstract.

7.2.3 Mapping external database information to ProMiner results

An important step in the adaptation procedure is the generation of database mapping and entity description files. The database mapping file contains database identifiers from external databases that are mapped to the corresponding gene/protein entry and ProMiner internal identifier. The database mapping file was created using the .map file generated during the compilation step as a base point, thus creating an enriched mapping file. The original file (.map file from ProMiner run) contained an internal ProMiner identifier, which was mapped to the given external database identifier and a synonym (an

example for an entry in ProMiner .map file BTC_000404:A1L595@SWISSPROT|KRT17 (see Figure 7.14)).

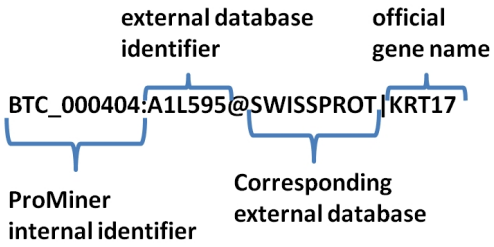


Figure 7.14: ProMiner .map file entry

As a part of this thesis, for generating .map files for cattle, pig and microRNA dictionaries external database mapping information were retrieved from databases and using perl scripts written for this purpose, the external database information was mapped onto the internal identifier (see Figure 7.15) for information enrichment of the found entities with external database information in SCAIView.

BTC_000404:A1L595@SWISSPROT|281889@ENTREZGENE|0005737 @GO|0005882@GO|0005198@GO| ENSBTAG00000006806@ENSEMBL| KRT17
an example for .map file entry with external database mapping.

A short description file was also created based on ProMiner generated .map file. The description file contains short description or officially recognized gene/protein name mapped to the internal identifier for the entity. The description file was used in SCAIView to give a brief description of the found entities. For curated dictionaries, the external database mappings and gene protein descriptions were retrieved from UniProt database, and Swissprot/Uniprot Interpro mappings were retrieved through Biomart service integrated to Ensembl database. Using perl scripting the retrieved external database mappings was mapped to ProMiner internal identifiers to create .map file. In the case of generated dictionary for cattle, the database mappings to Uniprot, Interpro, Gene Ontology annotations and Ensembl database were retrieved through Biomart integrated to Ensembl database and KEGG pathway mapping was retrieved from KEGG database. For SNP mapping, the xml files in which rs numbers were allocated to the respective Entrez Gene identifiers in each chromosome were retrieved from dbSNP ftp service and was parsed to get rs number to Entrez Gene mapping with a perl script written for the purpose. For pig generated dictionaries database mappings from Entrez Gene, Uniprot and KEGG was used along with functional gene annotation from Gene Ontology, since most other information on pig genes and proteins were not available at the time. For external database mapping of microRNAs dictionaries, the same methods used for external database mapping of gene/protein name dictionaries were used. The microRNAs were mapped to miRBase, currently existing microRNAs database and additional information

including the species in which the microRNA is present, the chromosome allocation of microRNAs, start and stop locations in the chromosomes and orientation of microRNA in the chromosome were added to the mapping file.

A mapping file created this way is used for external database information integration and result filtering. For example, for the gene KRT17 the GO identifiers found are 0005737, 0005882 and 0005198. SCAIView allows further search using these found results (search using results). On clicking the result filter icon in the entity page, the selected results are filtered. If the result filtering is done with the go id 0005737 in SCAIView, the filter search area shows the search query as "0005737@GO". When such a search is made, all the entities corresponding to the search term are selected and rest are filtered out.

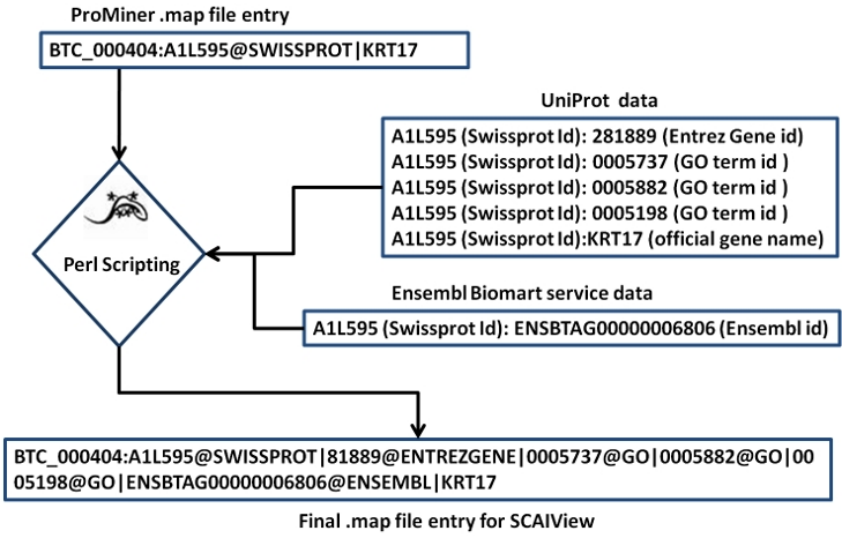


Figure 7.15: Workflow used for creating .map files for SCIAView, illustrated with an example

7.2.4 MicroRNA target analysis and mapping of computationally determined microRNA targets to SCAIView results

Since microrna dictionaries were created, the target analysis of these microRNAs and integration the obtained result to the SCAIView results were done for this thesis. For microRNA target analysis, miRanda algorithm was used (refer to Section 7.1.7). The microRNAs target sequences were retrieved from miRBase database and target 3' UTR sequences of the target organism genes were retrieved from Ensembl database in FASTA format. The two files were used as input files in miRanda algorithm. The output files containing the results given by the algorithm was further parsed to make a refined list of microRNAs targets. The parsing process was done using perl script. MicroRNA target analysis and integration of target Entrez Gene identifiers to microRNA dictionary mapping file was done to make querying for microRNAs and its targets to an automated

way. The current microRNAs target searches in texts involve searching with relevant keywords and manual scanning for microRNAs gene targets in texts. MicroRNA targets for pig microRNA in pig genome, cattle microRNA in cattle genome and human and mouse microRNA in cattle and pig genome were computationally determined using the algorithm. The reason for analysing cattle and pig gene targets for human and mouse microRNAs is that like orthologue genes, microRNAs also tend to be evolutionary conserved in their function and sequence, they follow similar nomenclature and scientists have explored the possibility of using orthologue microRNAs in different organisms through experimental validation. The 3' UTR sequences were initially mapped to their respective Ensembl gene identifiers, since the data was retrieved from Ensembl in FASTA format and header sections of the FASTA file was the corresponding Ensembl identifier. So for all the final hits identified, the corresponding Entrez gene identifiers were retrieved through Ensembl Biomart service. Similarly, the header sections of the microRNAs FASTA file had microRNA names, so these microRNA names were mapped to miRBase identifiers. Ensembl gene identifier to Entrez gene mapping, microRNA name to miRBase identifier mapping and a modified ProMiner .map file for microRNA were used as inputs for a perl script (written for this particular purpose), that used all the data to generate a final mapping file for microRNA dictionaries and was used for microRNA data enrichment in SCAIView.

The microRNA targets were mapped to the corresponding microRNAs to enable search using results. For example, from the microRNA bta-miR-2443, a user could easily navigate to one of the possible targets for the microRNA and can extract the external database information and functional gene annotations for the target gene.

7.2.5 Terminology analysis of cattle preimplantation period

Terminology analysis and integration of the terminology to SCAIView was done to facilitate a terminology based search. For this thesis, terminology analysis of cattle preimplantation period was done with the help of the researchers from the institute of animal science. The terms coined by the researchers were classified under four different headings. The terminology was designed to include all the major embryo structures, structural components and developmental processes that a preimplantation embryo undergoes, as well as all the major artificial embryo manipulation techniques and procedures that are used in preimplantation embryo manipulation. The terminology hierarchy was arranged into four major sections, 'Embryo stages', 'Embryo development', 'Sources' and 'Procedures'. The first section in the terminology hierarchy was 'Embryo stages', in this stage included terms describing different stages and morphological structures of an embryo during the preimplantation development. Embryo stages branch was subdivided into two sections 'Oocyte' and 'Embryo'. Although oocyte does not directly come under cattle preimplantation period, oocyte was included because of the major role that oocyte plays in preimplantation embryo development. Three subsections, 'cleavage stages', 'Morula' and 'Blastocyst' were included in 'Embryo' section. These subsections are different stages of preimplantation embryo development and each of them plays a key role in embryo development. The inner divisions and final nodes of these subsections include all the

major structures and structural components that form an embryo at the mentioned stage. The second major section was 'Embryo Development', which includes all the developmental processes that an embryo undergoes during preimplantation stage. This section had two major subdivisions, 'Morphological development' and 'Non morphological development'. The first subdivision incorporated all the preimplantation embryo developmental process which changed the morphology of the embryo, like cleavage division and embryo compaction. The second subdivision covered all the embryo developmental procedures that do not change the morphology of the embryo but plays a major role in preimplantation development, like embryo genome activation. It was found that some of the structure entities that need to be defined in the 'Embryo section' were already defined in 'Human developmental anatomy timed version[27]' ontology in the OBO foundry. So, all the 'Embryo structure' relevant found in 'Human developmental anatomy timed version' ontology were added to the terminology (for example: polar trophectoderm (id: EHDA:111) and mural trophectoderm (id: EHDA:59), different classification of trophectoderms). Similarly some of the entities in GO biological process ontology was found to be relevant to 'Embryo development' section of the terminology and were added to the respective terminology section (for example inner cell mass cell differentiation (GO:0001826) and inner cell mass cellular morphogenesis (GO:0001828)). The remaining two sections were aimed at the embryo manipulation techniques and procedures that were used. The section 'Sources' defined source of an embryo like its environment (in-vitro or in-vivo) and the cloning process used in the embryo generation. The last section, 'Procedures' accounts for the artificial techniques and natural procedures that are found in association with the embryo environment and cloning processes. For example, 'embryo transfer' is an entry in 'Procedures' that could be used in association with 'in vitro' in 'Environment' hierarchy as 'in vitro embryo transfer'.

The terminology so generated was converted into ProMiner dictionary format with a local internal identifier as a unique identifier instead of an external database identifier. The different spelling variations and acronyms of the terms in terminology are added to the dictionary and a ProMiner run was done on the previously mentioned, ∼470,000 abstracts that were used for ProMiner run of the gene protien dictionaries. The resulting prt file was also indexed with SCAIView to create an indexed of the terms. An XML file was generated (see Figure 7.16) following the hierarchies explained and with all the terms in the terminology dictionary. The XML file was integrated into SCAIView to allow terminology based search. The terminology was developed independently, not based on Mesh or UMLS terms since the cattle preimplantation domain was too specific, and the terms included in Mesh or UMLS are representing a broader concept.

Similar to result based search, ontology or terminolgy based search filters the entities and documents depending on the occurrence of entities along with the ontology or documents containing the term. An example query and the based results are explained in Section 8.3.

[27] http://www.obofoundry.org/cgi-bin/detail.cgi?id=human-dev-anat-staged last accessed 29 September 2009

```
- <tree id="run_6">
  - <item id="6_1" text="Embryo Stages">
    - <item id="6_2" text="Oocyte">
        <item id="6_3" pid="IMPL_000001" text="Immature oocyte" />
        <item id="6_4" pid="IMPL_000002" text="Mature oocyte" />
      </item>
    - <item id="6_5" text="Embryo">
      - <item id="6_6" pid="IMPL_000003" text="cleavage stages">
          <item id="6_7" pid="IMPL_000017" text="2 cell stage" />
          <item id="6_8" pid="IMPL_000018" text="4 cell stage" />
          <item id="6_9" pid="IMPL_000019" text="8 cell stage" />
          <item id="6_10" pid="IMPL_000020" text="16 cell stage" />
        </item>
      - <item id="6_11" pid="IMPL_000004" text="Morula">
          <item id="6_12" pid="IMPL_000021" text="noncompacted morula
          <item id="6_13" pid="IMPL_000022" text="Compact morula" />
        </item>
      - <item id="6_14" pid="IMPL_0000015" text="Blastocyst">
        - <item id="6_15" pid="IMPL_000032" text="Inner cell mass">
            <item id="6_16" pid="IMPL_000040" text="Epiblast" />
            <item id="6_17" pid="IMPL_000041" text="Hypoblast" />
          </item>
        - <item id="6_18" pid="IMPL_000042" text="Trophectoderm">
            <item id="6_19" pid="IMPL_000024" text="Trophoblast" />
            <item id="6_20" pid="IMPL_000043" text="Polar trophectoderm
            <item id="6_21" pid="IMPL_000044" text="Mural trophectoderm
          </item>
          <item id="6_22" pid="IMPL_000045" text="Blastocoel" />
        </item>
      </item>
    </item>
```

Figure 7.16: Terminology in XML format

7.2.6 Indexing

Indexing is done to improve data retrieval operations in a database table. SCAIView indexing was done by using the above mentioned PMIDs of the abstracts as source list for abstracts (texts) and the PRT files from ProMiner were used as source list for list of terms to be indexed. The indexing process was done with PRT files from gene and protein dictionaries generated for cattle and pig, microRNA dictionaries of cattle, human, mouse and pig, terminology dictionary, Gene Ontology descriptions and SNP descriptions that were created as a part of an earlier SCAIView project as a part of @neurIST project. For the final indexing procedure, the list of PubMed identifiers were given from the veterinary corpus[28]. So, instead of using the previously mentioned 470,000 abstracts for the indexing process, the process was done on approximately 1,599,000 abstracts, the total number of abstracts in the veterinary corpus.

A complete, generalized workflow adopted during this thesis is given in Figure 7.17.

[28]http://www.nlm.nih.gov/services/veterinarymed_details.html last accessed 10 November 2009

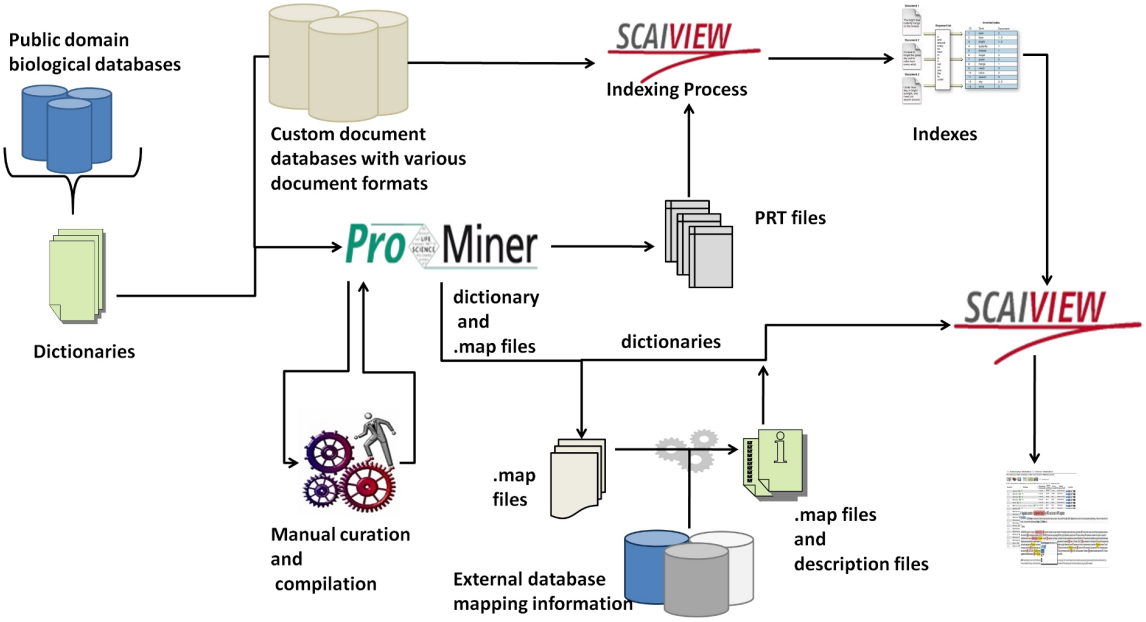


Figure 7.17: Generalized workflow for generation and integration of dictionaries into SCAIView

8 Results and discussion

This chapter describes the the final results obtained and further discussions on the results. The needed files such as the dictionaries, external database mapping files, description files and database indexes were generated as mentioned in the methodology section and were integrated into the SCAIView system, and test runs were made to test the system. Figure 8.1 shows the animal version home page and Figure 8.2 shows the entity result page.

Figure 8.1: Animal SCAIView homepage

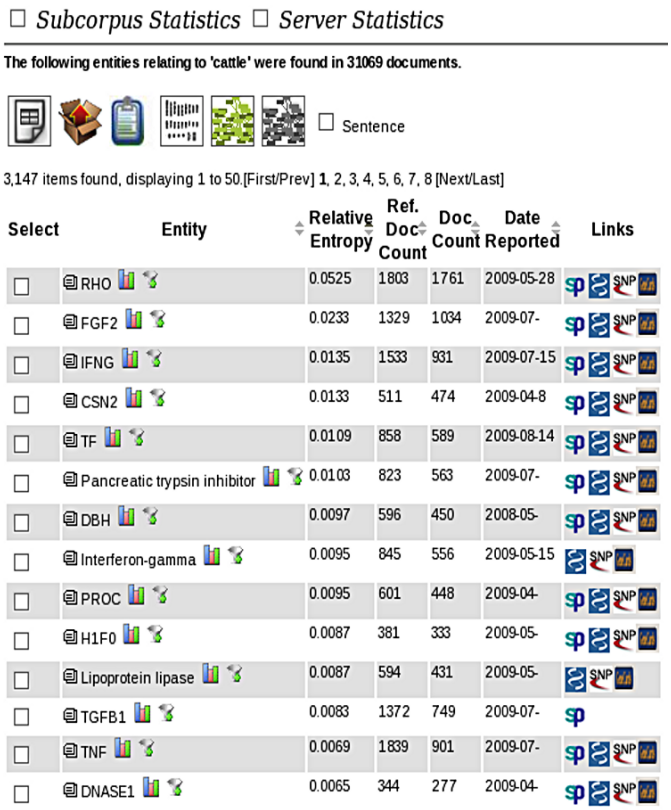


Figure 8.2: Animal SCAIView entity view

The performances of the gene and protein dictionaries were assessed, the interaction network given by SCAIView is analysed and Miranda algorithm is used to predict possible cattle microRNAs that target preimplantation genes. The results of these various performance analysis are explained in this section.

8.1 Performance analysis

The performance analysis of the system was done by calculating precision, recall and F score as explained in Section 3.5. The results are given below:

Table 8.1: Annotated corpora and number of extracted entities

Corpora	Number of abstracts	Number of abstracts annotated gene and protein entities	Extracted entities
Cattle general corpus	200	46	413
Cattle specific corpus	200	179	1846
Pig general corpus	200	64	326
Pig specific corpus	200	184	1626

Table 8.2: Annotated corpora and number spelling variants and unique entities

Corpora	Entities extracted	Spelling variants,case sensitive entities	unique Entities
Cattle general corpus	413	157	124
Cattle specific corpus	1,812	588	-
Pig general corpus	326	159	129
Pig specific corpus	1,626	545	-

The performance analysis was done on three sets of dictionaries:
Curated dictionaries: Cattle and pig gene/protein dictionaries created from Uniprot data of the organisms
Generated dictionaries: Dictionaries created from orthologue mapping of cattle and pig genes to human and mouse counterparts.
Combined dictionaries: For each organism the curated and generated dictionaries were combined into a single dictionary to analyze the performance.
In addition, the prt files obtained from the ProMiner run for curated and generated dictionaries of each organism were combined and the performance of the combined results were also assessed.

All the three sets of dictionaries were given as seed dictionaries in ProMiner as explained in section 7.1.11 and the the list of PubMed identifiers, which were used for annotation were given as base document for entity tagging. The generated PRT files were retrieved and analyzed, in addition, the ProMiner PRT files obtained from individual dictionaries, (curated and generated dictionaries) were combined and the result was also analyzed.

Since two separate dictionaries were created for a single organism, the overlap between the two dictionaries was identified. For this purpose, the number of test abstracts in which gene and protein mentions were found by both the dictionaries was identified. Similarly, the number of test abstracts in which gene protein mentions uniquely identified by each of the dictionaries was also identified. The results were analysed for both corpora,

the general corpus with documents covering a large domain and the specific corpus, the corpus that is biased towards cattle preimplantation genomis and pig meat quality and genetics.

Table 8.3: Dictionaries and entities found in ProMiner run for general corpus

Corpora	Dictionary	Number of abstracts with entities found	Extracted entities
Cattle general corpus	specific dictionary	56	355
	orthologue dictionary	55	395
	combined dictionary	44	197
Cattle specific corpus	specific dictionary	166	1585
	orthologue dictionary	173	1592
	combined dictionary	123	805
Pig general corpus	specific dictionary	50	187
	orthologue dictionary	41	176
	combined dictionary	39	113
Pig specific corpus	specific dictionary	141	1130
	orthologue dictionary	155	1379
	combined dictionary	113	720

The prt files returned after the test run were also analyzed to find the number of abstracts in which entities are identified uniquely by each dictionary, and the number of abstracts in which entities were identified by both dictionaries in common. The results are given in the table below:

Table 8.4: Table showing overlapping and unique abstracts

Organism	Copora	Dictionary	Number of abstracts with entities found	Common for both dictionaries	Unique for each dictionary
Cattle	general corpus	specific dictionary	56	44	12
		Orthologue dictionary	55	44	11
	specific corpus	specific dictionary	166	158	8
		Orthologue dictionary	173	158	15
Pig	general corpus	specific dictionary	50	29	21
		Orthologue dictionary	41	29	12
	specific corpus	specific dictionary	141	117	24
		Orthologue dictionary	155	117	38

From the extracted annotations and prt files from ProMiner, number of true positives, false positives and false negatives were calculated using a perl script and the results are given below:

Table 8.5: Performance analysis: True positive, false positive and false negative values for dictionaries on general and test corpora

Organism Corpora	Dictionary	True positive	False positive	False negative
Cattle general corpus	specific dictionary	244	111	169
	orthologue dictionary	264	131	149
	dictionaries merged	110	87	303
	results combined	308	170	105
Cattle specific corpus	specific dictionary	1,376	207	436
	orthologue dictionary	1379	211	433
	dictionaries merged	627	166	1,185
	results combined	1,583	298	229
Pig general corpus	specific dictionary	143	44	183
	orthologue dictionary	130	46	196
	dictionaries merged	62	51	264
	results combined	175	76	151
Pig specific corpus	specific dictionary	985	142	641
	orthologue dictionary	1,201	178	425
	dictionaries merged	586	123	1,040
	results combined	1,320	233	306

From the extracted entities precision, recall and F score were calculated for all the three sets of dictionaries and the result obtained by merging the dictionaries. The table given below shows the results (rounded to two digits):

Table 8.6: Performance analysis Precision, Recall and F_1 score for dictionaries on general and test corpora

Organism and Corpora	Dictionary	Precision	Recall	F-Score
Cattle general corpus	specific dictionary	0.69	0.59	0.64
	orthologue dictionary	0.67	0.64	0.65
	dictionaries merged	0.56	0.27	0.36
	results combined	0.64	0.74	0.69
Cattle specific corpus	specific dictionary	0.87	0.76	0.81
	orthologue dictionary	0.87	0.76	0.81
	dictionaries merged	0.79	0.35	0.48
	results combined	0.84	0.87	0.86
Pig general corpus	specific dictionary	0.76	0.44	0.56
	orthologue dictionary	0.74	0.40	0.52
	dictionaries merged	0.59	0.19	0.28
	results combined	0.76	0.44	0.56
Pig specific corpus	specific dictionary	0.87	0.61	0.72
	orthologue dictionary	0.87	0.74	0.80
	dictionaries merged	0.83	0.36	0.50
	results combined	0.85	0.81	0.83

The prt files were obtained were also analyzed for the spelling variants and unique genes identified. The analysis was done by comparing the prt file results to the parsed annotation data and the results are shown in Table 8.10.

Table 8.7: Performance analysis: Dictionaries and number of spelling variants and unique entities found for each corpora

Organism	Corpus	Dictionary	Spelling and case variants found	Unique entities
Cattle	general corpus	specific dictionary	120	102
		orthologue dictionary	128	107
		results combined	166	
		combined dictionary	79	69
	specific corpus	specific dictionary	403	274
		orthologue dictionary	397	275
		combined dictionary	231	177
		results combined	482	
Pig	general corpus	specific dictionary	83	70
		orthologue dictionary	74	57
		combined dictionary	57	46
		results combined	115	
	Specific corpus	specific dictionary	284	209
		orthologue dictionary	345	219
		combined dictionary	183	146
		results combined	388	

The reason for the low performance of the combined dictionaries were that ‘the ambiguity filter functionality‘ in ProMiner was activated during the runs, which filtered out multiple occurrences of synonyms. So the performance of the dictionaries were analysed after deactivating the ambiguity filter functionality.

Table 8.8: Performance analysis: True positive, false positive and false negative values for dictionaries on general and test corpora with ambiguity filter removed

Organism and Corpora	Dictionary	True positive	False positive	False negative
Cattle general corpus	specific dictionary	274	183	139
	orthologue dictionary	272	209	141
	dictionaries merged	304	208	109
	results combined	312	289	101
Cattle specific corpus	specific dictionary	1,403	285	409
	orthologue dictionary	1415	408	397
	dictionaries merged	1,575	534	237
Pig general corpus	specific dictionary	191	70	135
	orthologue dictionary	147	78	179
	dictionaries merged	200	116	126
Pig specific corpus	specific dictionary	1,198	222	428
	orthologuedictionary	1,222	192	404
	dictionaries merged	1,333	255	293

Since a difference was found in the number of true positives, false positives and false negatives especially in case of combined dictionaries, precision recall and F-score of the dictionaries were analyzed.

Table 8.9: Performance analysis: Precision, Recall and F_1 score for dictionaries on general and test corpora with ambiguity filter removed

Organism and corpora	Dictionary	Precision	Recall	F_1 Score
Cattle general corpus	specific dictionary	0.52	0.76	0.62
	orthologue dictionary	0.57	0.66	0.61
	dictionaries merged	0.51	0.74	0.60
	results combined	0.51	0.76	0.62
Cattle specific corpus	specific dictionary	0.83	0.77	0.80
	orthologue dictionary	0.77	0.78	0.77
	dictionaries merged	0.74	0.86	0.80
	results combined	0.75	0.87	0.80
Pig general corpus	specific dictionary	0.73	0.59	0.65
	orthologue dictionary	0.65	0.45	0.53
	dictionaries merged	0.63	0.61	0.62
	results combined	0.63	0.61	0.62
Pig specific corpus	specific dictionary	0.84	0.74	0.79
	orthologue dictionary	0.86	0.75	0.80
	dictionaries merged	0.84	0.82	0.83
	results combined	0.83	0.83	0.83

The statistics from the indexing procedures were also analysed to find the number of entities identified by the dictionaries and the number of documents in which entities were found. The statistics were analyzed for all the gene/protein dictionaries and miRNA dictionaries along with cattle preimplantation terminology dictionary. The statistics are shown in the table below:

Table 8.10: Indexing statistics

Organism	dictionary	Corpora size	Number of entities found	Relative entity count (entities per document)	Number of documents with entities found	Unique entities found
Cattle	specific dictionary	1,599,024	964,909	3.87	249,596	5,665
	orthologue dictionary	1,599,024	1,088,870	4.13	263,850	7,858
	preimplantation terminology	1,599,024	361,827	2.0	179,663	35
	miRNA dictionary	1,599,024	456	4.19	109	93
Pig	specific dictionary	1,599,024	820,811	3.8	215,007	2,808
	orthologue dictionary	1,599,024	799,758	3.9	205,044	3266
	miRNA dictionary	1,599,024	247	3.3	76	37
Human	miRNA dictionary	1,599,024	445	4.3	102	87
Mouse	miRNA dictionary	1,599,024	449	4.4	101	89

8.2 Analysis of interaction networks

SCAIView provides two kinds of protein-protein interaction networks: from documents using co-occurrence of proteins and from BIANA, from which experimentally confirmed protein-protein interaction networks for candidate entities can be retrieved. The co-occurrence network is analyzed here and experimental interaction data exported from SCAIView are visualized. The co-occurrence data was retrieved on abstract basis, not on sentence basis. For the analysis of cattle preimplantation network, protein protein interaction network obtained from SCAIView through co-occurrence was exported and visualized in Cytoscape (see Figure 8.3). The networks obtained by using cattle specific dictionary and cattle orthologue dictionary were compared to find the difference and it was noticed that the networks were identical except for a few gene/protein entries. Some of the entries from the cattle specific dictionary include: SLC38A4, LRRCC1, cytokeratin 19, HBEGF, LGALS3, IGFBP3, CDX2, MKRN3 and some from the orthologue dictionary include: PTGFR, EGF, MBD3, DICER1, Hsp70, IFNT1, BMP15.

8.2.1 Protein networks from co-occurrence

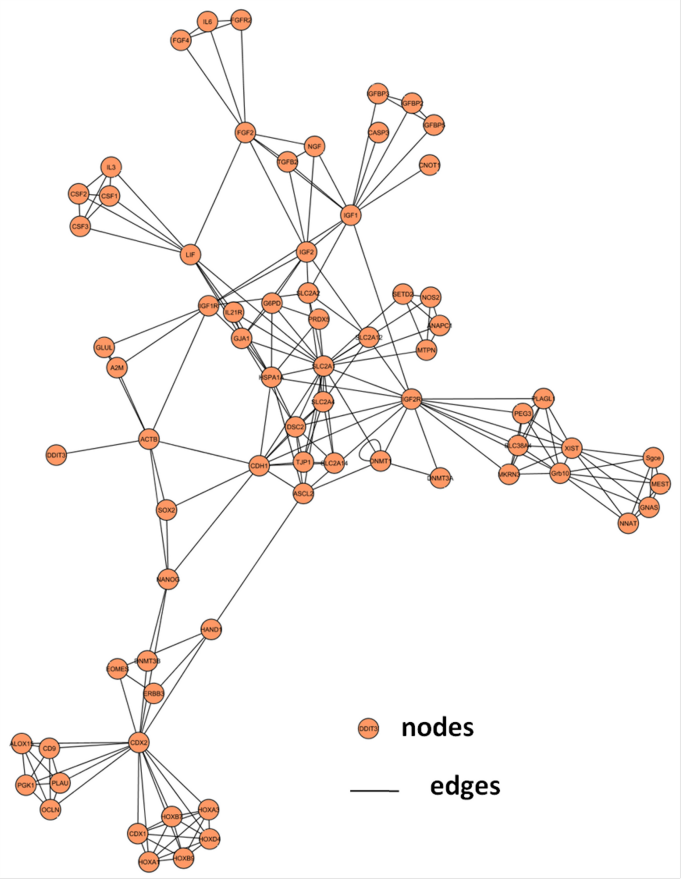


Figure 8.3: Cattle preimplantation interactions from SCAIView (co-occurrence data)

The interactions obtained from SCIAView needed to be confirmed by using the experimental data. But due to the lack of enough experimental data from cattle, experimental data from human and mouse were also considered. The gene/protein names obtained from SCAIView data were already normalized to Entrez gene and Uniprot entries, so these gene names were used as seed proteins and genes to extract interaction data from String database. The interactions that were found both in the SCAIView network and network obtained from String database were confirmed as positive interactions and others were considered as false interactions. Around 300 gene/protein names were given as seed gene/protein entries in String database and the resulting interactions data for cattle, human and mouse were retrieved from the database. From the analysis it was found that only a small fraction of the interactions from SCAIView interaction network was actually confirmed (see Figure 8.4).

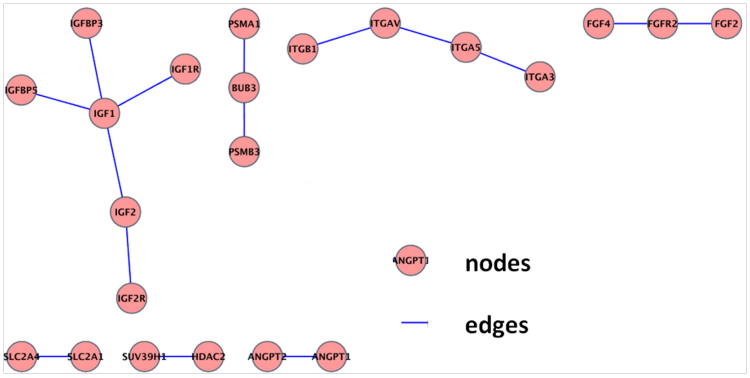


Figure 8.4: Confirmed interactions

Similar to cattle network preimplantation network analysis, gene/protein interaction network for pig meat quality were also analyzed, but meat quality cannot be constrained to a particular stage in growth and development like preimplantation period, so gene/protein interaction network for major gene effects having a direct influence on meat quality were analyzed. The major gene effects are the sex chromosome effect, stress gene effect and napole effect. The SCAIView network obtained for sex chromosome effect and napole effect for both pig specific and orthologue dictionary were negligible (with only 2 to 4 interaction partners), so both of the networks were discarded and interaction network for stress gene effect (from search query “pig AND napole effect“) (see Figure 8.5) was taken into consideration. As a first step the stress gene effect network obtained from pig specific dictionary and pig orthologue dictionary were compared and found that there is a noticeable difference in the interaction network given by two dictionaries. There was a lack of any experimental data for validating pig interaction network. The approach of combining human and mouse protein interaction data and analyzing the result also produced negligible result. So validation results for pig interaction networks are avoided as there is far less experimental data on porcine protein interaction networks.

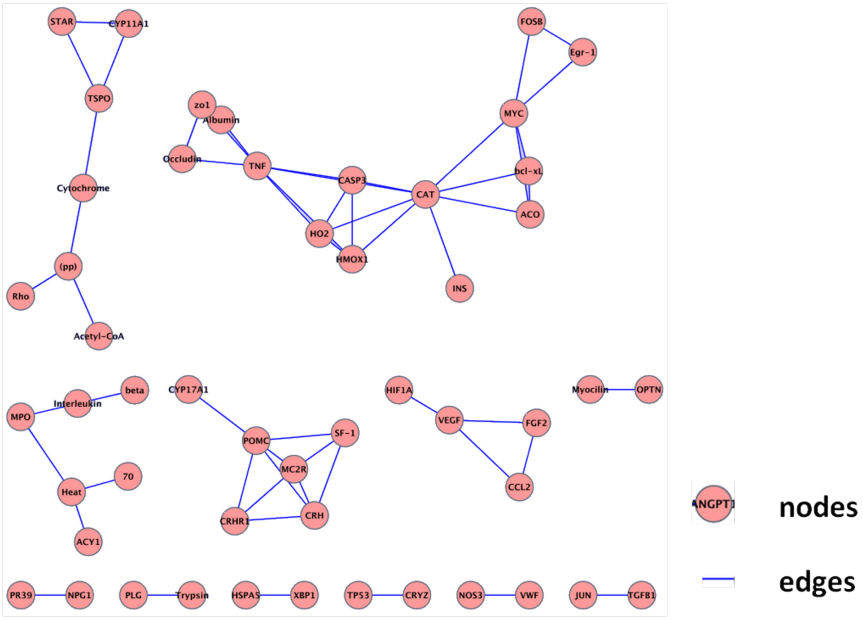


Figure 8.5: Pig stress gene effect SCAIView interactions

8.2.2 Protein networks from BIANA

The experimentally confirmed protein network data from SCAIView, obtained from BIANA in XGMML format[1] were retrieved and visualized in Cytoscape (see Figure 8.6). For the analysis, the search query "cattle" was used to retrieve a large list of gene/protien entities in SCAIView and three gene entries were selected from a list of gene entries, which were having experimental data in BIANA databases. The seed genes selected were VIM (Vimentin), GFAP (glial fibrillary acidic protein) and HSPA8 (heat shock 70kDa protein 8).

[1]http://www.cs.rpi.edu/~puninj/XGMML last accessed 8 November 2009

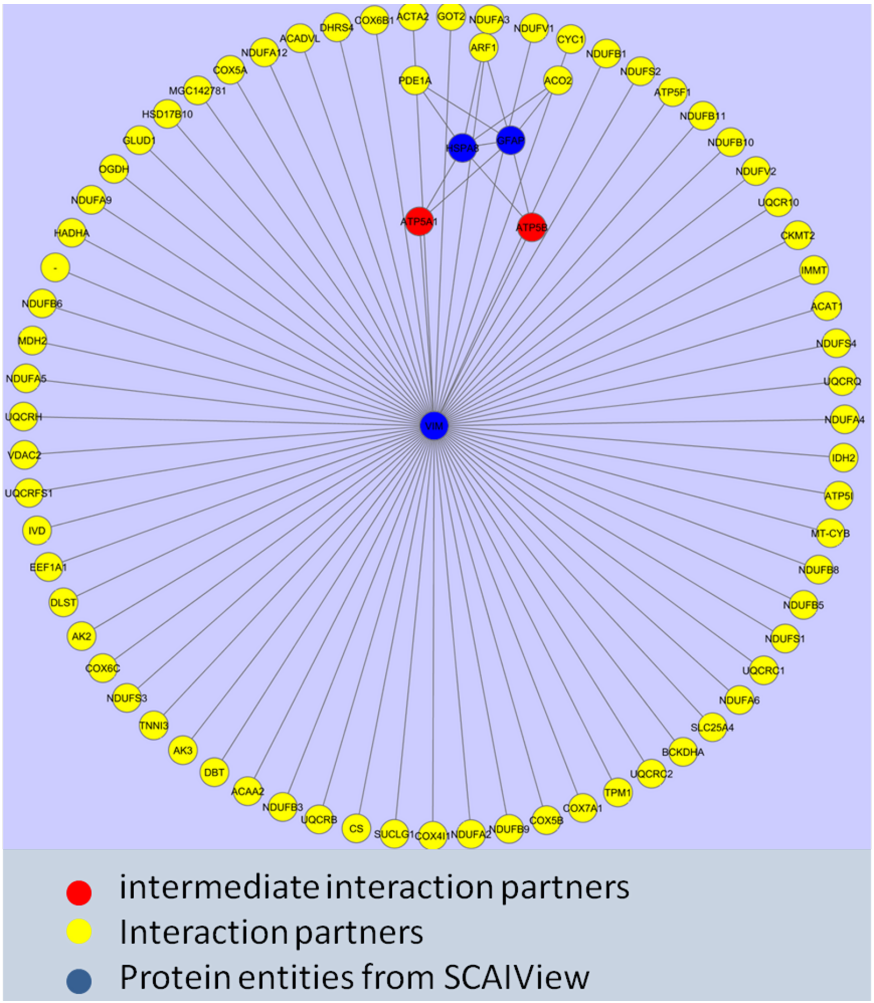


Figure 8.6: Experimentally confirmed protein interactions from BIANA

Vimentin (Entrez Gene id:280955) is an intermediate filament family protein, which exist as a dynamic structure. The functions of the protein include maintainig cell flexibility and cell integrity. The interactions of the vimentin were experimentally discovered for human and mouse genome and it was found that vimentin interacts with glial fibrillary acidic protein (GFAP, Entrez Gene id:281189), another intermediate filament family protein (Intact id: EBI-755266) (Rual et al., 2005). From the confirmed interactions, vimentin was also found to interact with HSPA8 (Entrez Gene id:281831), a protein that is involved in the thermotolerance process and stabilizes intermediate filaments through direct or indirect binding (Lee and Lai, 1995). For the interaction between GFAP and HSPA8, although Biomyn database shows an interaction[2], further proofs for the interaction were not available from the IntAct database link provided. Similar is the case with interactions between ATP5A1 and VIM and VIM and ATP5B.

[2]http://www.biomyn.de/index.php?mid=P11142&ispc=UniProtKB last accessed 12 November 2009

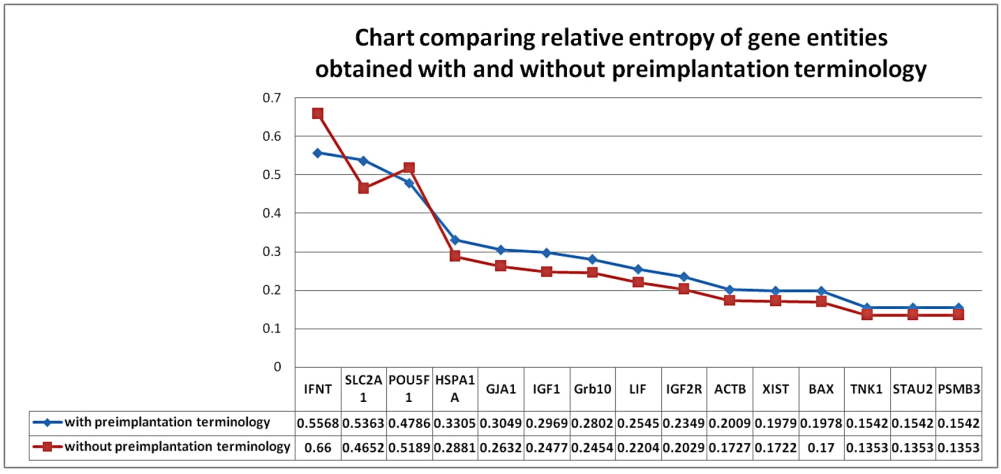
8.3 Analysis of preimplantation terminology

For the analysis of the cattle preimplantation terminology, a test search was performed in SCAIView without using the preimplantation terminology and the same query was repeated after selecting and using the preimplantation terminology, to include the terminology as well in the search. First of all, a simple search was made with the query "cattle preimplantation" in SCAIView without using the preimplantation terminology. For the search, it was found that a total of 55 documents were retrieved and 108 gene entities were identified by SCAIView. The same search was repeated after selecting the preimplantation terminology and it was seen that the number of documents were reduced to 49 and the number of gene entities dropped down to 98. For further analysis the relative entropy and document count of the first 15 hits from the search using preimplantation terminology was compared with that of the a search without using preimplantation terminology.

Table 8.11: Positions of gene entities in search with using preimplantation terminology compared to that of search without using preimplantation terminology

Gene	Position in search with terminology	Position in search without terminology
IFNT	1	1
SLC2A1	2	3
POU5F1	3	2
HSPA1A	4	4
GJA1	5	5
IGF1	6	6
Grb10	7	7
LIF	8	8
IGF2R	9	10
ACTB	10	15
XIST	11	15
BAX	12	23
TNK1	13	24
STAU2	14	25
PSMB3	15	26

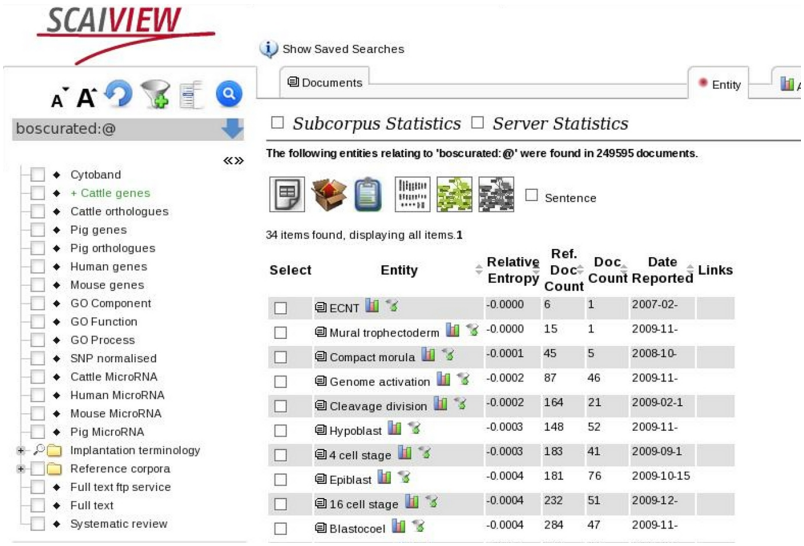
Since the found entities were sorted according to the relative entropy score in SCAIView, the change in the top ranking hits were due to the change in relative entropy score (refer to Section 7.1.12). A graph was plotted with the relative entropy score of the above hits (see Figure 8.7). On comparison it was found that except for two gene entities, all the others showed a higher relative entropy score.



	IFNT	SLC2A1	POU5F1	HSPA1A	GJA1	IGF1	Grb10	LIF	IGF2R	ACTB	XIST	BAX	TNK1	STAU2	PSMB3
with preimplantation terminology	0.5568	0.5363	0.4786	0.3305	0.3049	0.2969	0.2802	0.2545	0.2349	0.2009	0.1979	0.1978	0.1542	0.1542	0.1542
without preimplantation terminology	0.66	0.4652	0.5189	0.2881	0.2632	0.2477	0.2454	0.2204	0.2029	0.1727	0.1722	0.17	0.1353	0.1353	0.1353

Figure 8.7: Chart comparing relative entropies of gene entities from search with preimplantation terminology and with preimplantation terminology

A further analysis was done to find the specific terminologies entities that were found along with cattle genes. For this purposes, a SCAIView search was made for cattle preimplantation terminology with the search query "boscurated:@", which is the name of the cattle specific gene dictionary prt file used in indexing process. The search was made for cattle preimplantation terminology, but only using cattle genes. It was seen that the term ECNT, (embryonic cell nuclear transfer) was the most specific preimplantation terminology associated with cattle specific dictionary (see Figure 8.8).



Select	Entity	Relative Entropy	Ref. Doc Count	Doc Count	Date Reported	Links
	ECNT	-0.0000	6	1	2007-02-	
	Mural trophectoderm	-0.0000	15	1	2009-11-	
	Compact morula	-0.0001	45	5	2008-10-	
	Genome activation	-0.0002	87	46	2009-11-	
	Cleavage division	-0.0002	164	21	2009-02-1	
	Hypoblast	-0.0003	148	52	2009-11-	
	4 cell stage	-0.0003	183	41	2009-09-1	
	Epiblast	-0.0004	181	76	2009-10-15	
	16 cell stage	-0.0004	232	51	2009-12-	
	Blastocoel	-0.0004	284	47	2009-11-	

Figure 8.8: Figure showing the search strategy and the specific preimplantation terms associted with cattle specific dictionary.

8.4 Prediction of cattle miRNAs targeting preimplantation genes

This section describes cattle microRNAs that are found to target cattle preimplantation genes through computational prediction. This analysis was done in order to predict cattle microRNAs that target major cattle preimplantation genes. Miranda algorithm was used to find candidate microRNAs that target cattle preimplantation gene. Miranda algorithm and the methodology followed are explained in the previous chapter. The result file from the Miranda run on cattle 3' UTR sequences is used here. The output file is parsed and cattle miRNA and targeting genes are extracted from the file and the top scoring microRNAs that are found to target cattle preimplantation genes are given in the result table. The table do not contain all the microRNAs that are found to target cattle preimplantation genes, but contains some of the top scoring microRNAs that are found to target some of the cattle preimplantation genes.

Table 8.12: Predicted cattle preimplantation microRNAs

Gene	Ensembl Gene id	Number of predicted microRNAs	Top scoring microRNAs
beta-catenin	ENSBTAG00000016420	4	bta-miR-2349, bta-miR-340 bta-miR-409 , bta-miR-664
CX43	ENSBTAG00000001835	47	bta-miR-2328, bta-miR-148a bta-miR-30e-5p, bta-miR-2305 bta-miR-2466-3p
FGF-2	ENSBTAG00000005691	42	bta-miR-34a, bta-miR-224 bta-miR-484, bta-miR-449a bta-miR-449b
GLUT-1	ENSBTAG00000018647	38	bta-miR-2338, bta-miR-2376 bta-miR-2377, bta-miR-2338 bta-miR-2306
IGF-1	ENSBTAG00000011082	13	bta-miR-138, bta-miR-329b bta-miR-370, bta-miR-658 bta-miR-138
IGF2R	ENSBTAG00000002402	19	bta-miR-615, bta-miR-2388 bta-miR-2467, bta-miR-2433 bta-miR-197
IGFBP3	ENSBTAG00000014541	31	bta-miR-2334, bta-miR-486 bta-miR-2322, bta-miR-2373 bta-miR-320
IGFBP5	ENSBTAG00000007062	26	bta-miR-2428, bta-miR-2305 bta-miR-2412, bta-miR-2309 bta-miR-2392
LIF	ENSBTAG00000007424	81	bta-miR-2466-3p, bta-miR-504 bta-miR-2343 bta-miR-2441, bta-miR-2447
NGF	ENSBTAG00000007446	3	bta-miR-2454, bta-miR-423-5p bta-miR-296
OCT4	ENSBTAG00000021111	30	bta-miR-412, bta-miR-145 bta-miR-2394, bta-miR-2326 bta-miR-346
PAFr	ENSBTAG00000027051	46	bta-miR-342, bta-miR-2349 bta-miR-2426, bta-miR-423-5p bta-miR-2428
PDGFA	ENSBTAG00000014541	31	bta-miR-2334, bta-miR-486 bta-miR-2322, bta-miR-2373 bta-miR-320
TNFA	ENSBTAG00000025471	12	bta-miR-2324, bta-miR-1584 bta-miR-2418, bta-miR-2382 bta-miR-2363
TGFB	ENSBTAG00000005359	1	bta-miR-677

During the analysis it was also noticed that almost all of the microRNAs that are found to target preimplantation genes target the gene IGF-II, suggesting a global expression of the gene. For the second part of the analysis cattle preimplantation genes and the targeting microRNA results were combined together and a perl script was used to cluster the results of each microRNAs together. This step was done to predict the target

specificity of the identified microRNAs on preimplantation genes. During the analysis it was found that certain microRNAs have specific targeting nature. During the analysis certain gene clusters were identified were a single microRNA was targeting a number of preimplantation genes.

Table 8.13: Micrornas and preimplantation gene targets

Microrna	Preimplantation gene targets found	Target gene Ensemble id	Gene name
bta-miR-2334	4	ENSBTAG00000014541	PDGFA
		ENSBTAG00000005691	FGF2
		ENSBTAG00000013066	IGF-II
		ENSBTAG00000021111	OCT4
bta-miR-2373	4	ENSBTAG00000014541	PDGFA
		ENSBTAG00000007424	bLIF
		ENSBTAG00000013066	IGF-II
		ENSBTAG00000021111	OCT4
bta-miR-615	5	ENSBTAG00000013066	IGF-II
		ENSBTAG00000027051	PAFr
		ENSBTAG00000002402	IGF2R
		ENSBTAG00000007424	bLIF
		ENSBTAG00000011082	IGF-1
bta-miR-2295	5	ENSBTAG00000007062	IGFBP5
		ENSBTAG00000013066	IGF-II
		ENSBTAG00000001835	Cx43
		ENSBTAG00000018647	SLC2A11
		ENSBTAG00000021111	OCT4
bta-miR-2443	5	ENSBTAG00000013066	IGF-II
		ENSBTAG00000002402	IGF2R
		ENSBTAG00000001835	Cx43
		ENSBTAG00000018647	SLC2A11
		ENSBTAG00000007062	IGFBP5
bta-miR-2324	5	ENSBTAG00000025471	TNFA
		ENSBTAG00000013066	IGF-II
		ENSBTAG00000007062	IGFBP5
		ENSBTAG00000007424	bLIF
		ENSBTAG00000018647	SLC2A11
bta-miR-2382	6	ENSBTAG00000025471	TNFA
		ENSBTAG00000007424	bLIF
		ENSBTAG00000014541	PDGFA
		ENSBTAG00000013066	IGF-II
		ENSBTAG00000018647	SLC2A11
		ENSBTAG00000001835	Cx43
bta-miR-1343	5	ENSBTAG00000013066	IGF-II
		ENSBTAG00000027051	PAFr
		ENSBTAG00000013066	IGF-II
		ENSBTAG00000007424	bLIF
		ENSBTAG00000018647	SLC2A11
		ENSBTAG00000002402	IGF2R

The prediction process gave a list of genes that are targeted by each cluster and it was found that the clusters target similar genes. So the sequence of mature miRNA sequences were analysed, but it was found that these microRNAs do not share sequence similarity. A characteristic feature of animal microRNAs is that they are approximately complementary to their targets (Ambros, 2004), which also leads to the hypothesis that microRNAs targeting the same gene need not be absolutely identical.Further experimental data is needed to validate these microRNAs and their targets.

8.5 microRNA targets from Text mining

The data from target prediction of cattle microRNA was used in augmenting microRNA information. SACIView can be used to search for a microRNA and its possible target. For example, the search query "cattle" for cattle microRNAs in SCAIView retrieves 6 microRNAs with target mapped. The microRNAs and their target Entez Gene ids are given in the Table 8.14.

Table 8.14: Micrornas and preimplantation gene targets

Cattle microRNA	Possible targets
bta-mir-181a	100125264
bta-mir-338	280968
bta-mir-10a	505436
bta-mir-30c	616179 505853 538100
bta-mir-26b	50583 618648 505757
bta-mir-222	535099 541294 100137795

From microRNA data, the user can linkout to the possible gene target and can extract the information relating to the targets.

8.6 Discussion

The results of performance and network analysis and predicted cattle preimplantation microRNAs are discussed in this section.

8.6.1 Performance analysis

From the result section it was clear that organism specific dictionary and orthologue dictionary had a high degree of overlap because of the sequence similarity between the

target and model organisms. It was clearly noticeable in case of cattle dictionaries, where there was a high degree of overlap than pig dictionaries (refer to Figure 8.9 and Figure 8.10). The reason for this high degree of overlap in cattle dictionaries than that of pig dictionaries can be related to the status of the cattle and pig genome sequencing project, where cattle sequencing projects are ahead of pig genome sequencing projects.

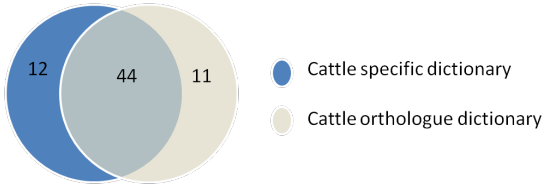


Figure 8.9: Overlap in cattle test abstracts

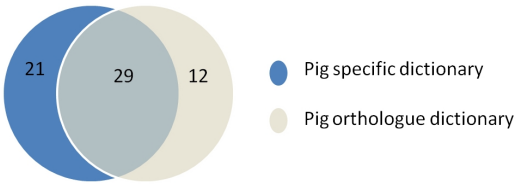


Figure 8.10: Overlap in pig test abstracts

The large overlap in the number of test abstracts in which gene/protein mentions were identified by both the dictionaries indicate the similarity of gene/protein annotations in the target organisms (cattle, pig), and the model organisms (human and mouse). The abstracts in which gene/protein mentions were identified only by the organism specific dictionary can contain gene/protein entries specific to the organism or gene protein annotations different from the model organisms (for example, the gene **CCNB1** could be mentioned as **bCCNB1** to denote bovine CCNB1). Uniprot and Entrez Gene entries of cattle genes and proteins showed that cattle and pig genome annotations followed human gene annotation schema and representation formats, where the gene names are represented in capital letters. Prostaglandin D2 synthase entry for cattle, human and mouse can be used to illustrate this fact (Entrez Gene ids: 19215 [*Mus musculus*], 286858 [*Bos taurus*], 5730 [*Homo sapiens*]) , where it can be seen that for both cattle and pig entries, the gene name is represented entirely in capital letters where as in case of mouse only the first letter of the gene name is in capital letter and the rest is in small letters. On comparing the cattle and human Entrez Gene entry of the gene, it could be noted that the most of the synonyms for the gene (for example like PGD2, PGDS and PGDS2) which are represented as ‘ Other Aliases ‘ for human gene are missing in the cattle entry and an Entrez Gene search with query ‘PGD2 AND bos taurus‘ returned no result. Several instances can be pointed out, where the synonyms of a gene/protein entry, which are not represented in public database gene/protein entries are used in scientific texts. The usefulness of orthologue dictionaries can be pointed out here. For example, the abstract “ Discovery of eight novel divergent homologs expressed in cattle placenta “ (PMID 16554549) mentions about homologs identified in cattle placenta including PRP11. Entrez Gene search for ‘PRP11‘ in cattle gave the gene SF3A3 (Entrez Gene id 523250) as the result, for which PRP 11 is a synonym, but not represented in the Entrez Gene entry. Uniprot search for cattle ‘PRP11‘, gave no result, which makes the representation of the gene in the abstract ambiguous to the user. So, a dictionary made using cattle gene and protein information would not identify PRP11 as a gene/protein entry. But, Homologene[3] links from the Entrez Gene id 523250 shows that human SF3A2 gene (Entrez Gene id

[3]http://www.ncbi.nlm.nih.gov/homologene last accessed 21 October 2009

8175) as an orthologue of the cattle gene, and the synonym 'PRP11' is represented as a synonym for the human SF3A2 gene. In cattle orthologue dictionary, human SF3A2 gene was mapped as an orthologue entry to cattle SF3A2 gene and synonyms from human SF3A2 gene were used as synonyms for cattle SF3A2 gene during the generation of cattle orthologue dictionary. So PRP11 gene mention would be identified by cattle orthologue dictionary as a synonym for cattle SF3A2 gene. The absence of such gene annotations in cattle gene entries can be a reason why certain gene/protein mentions in test abstracts were identified only by cattle and pig orthologue dictionaries.

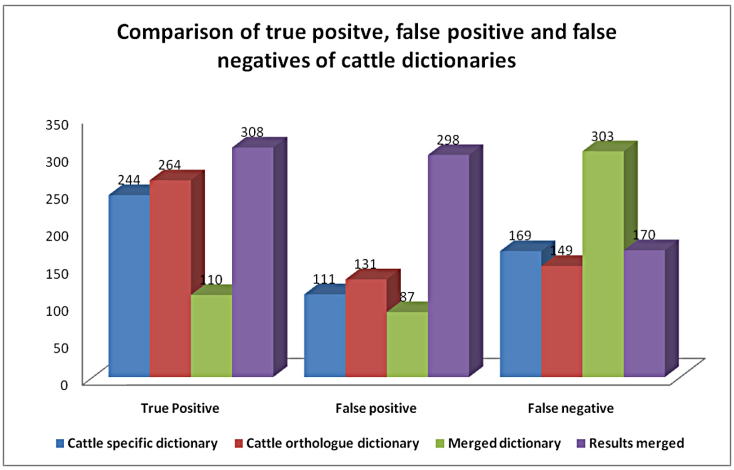


Figure 8.11: Cattle true positive false positive and false negative chart

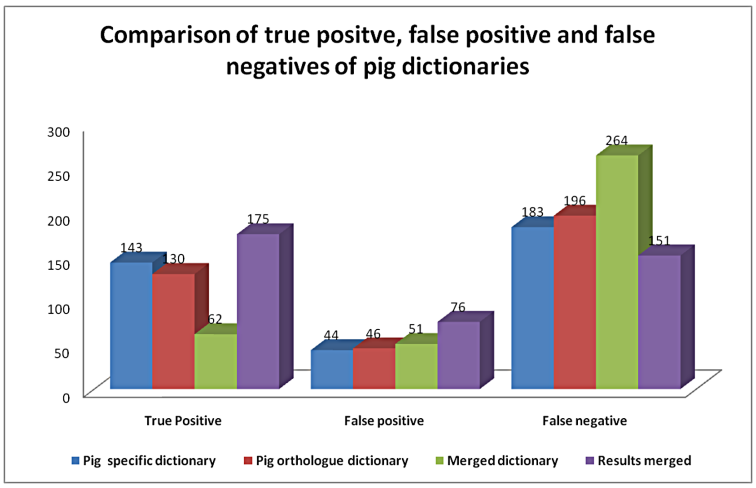


Figure 8.12: Pig true positive false positive and false negative chart

The comparison of true positives and false negatives of organism specific dictionary with that of organism orthologue dictionary reveals the quality and extent of gene/protein

annotations in test organisms cattle and pig. For cattle dictionaries (refer to Figure 8.11) it can be seen that the difference between true positives identified by cattle orthologue dictionary and cattle specific dictionary is only small, which could suggest the quality and extent of cattle genome annotation. For pig dictionaries (refer to Figure 8.12) it can be seen that there is a large difference in the true positives identified by pig orthologue dictionary and pig specific dictionary. The graphs also shows low true positive values, when organism specific and orthologue dictionaries were combined. The reason for low true positive values for merged dictionaries is ambiguity filter in ProMiner, which in its default setting removes all the synonyms with multiple occurrences in the same dictionary. For the test run, the ambiguity filter was in its default setting and removed all the synonyms with multiple occurrences. So, the true positive values from the merged dictionaries indicate synonyms in both organism specific or orthologue dictionary which does not occur in the other dictionary and found in the test abstracts and these synonyms can include organism specific synonyms from organism specific (curated) dictionary and synonyms found only in abstracts and gene/protein entries of model organism and not in target organism gene/protein entries in databases. Another reason can be the difference in manual curation steps, where a spelling variant of a specific gene/protein entry was added in a dictionary and missed in the other. For pig dictionaries, the performance scores are low and this can be as a result of less number of pig gene protein entities in the dictionaries. The number of curated gene protein entries in public database during the generation of dictionaries is the major factor behind the performance scores.

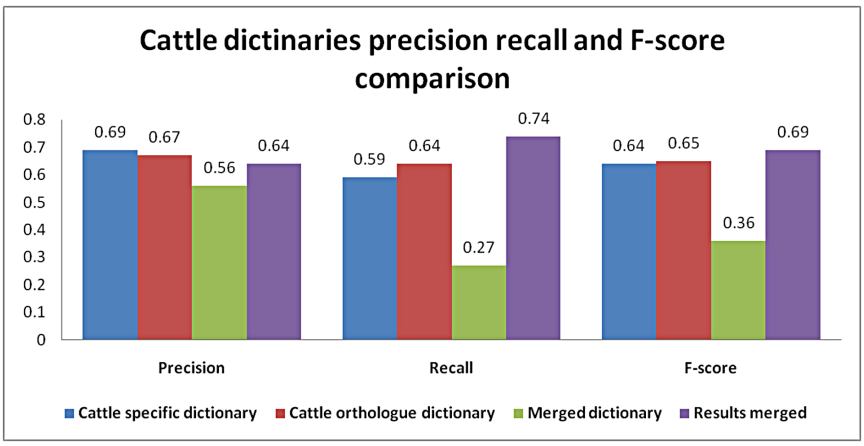


Figure 8.13: Cattle precision, recall and F-score chart

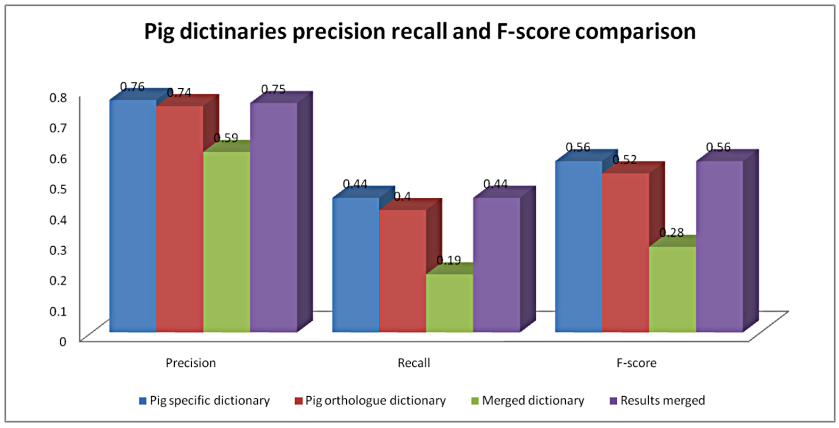


Figure 8.14: Pig precision, recall and F-score chart

The similarity in precision, recall and F scores of cattle dictionaries are as a result of similar true positive, false positive and false negatives values for the dictionaries in the test abstract (refer to Figure 8.13 and Figure 8.14). For the merged results (PRT files from ProMiner run of organism specific and orthologue dictionaries merged) it can be seen that the precision of the system was reduced by a small value and there was increase on the recall of the system. Since the prt files of both dictionaries were merged, the increase in recall is a result of extended coverage of the system, where the relevant synonyms from the test abstract that are absent in the organism specific dictionary are identified organism orthologue dictionary and those missed by organism orthologue dictionary are identified by organism specific dictionary.

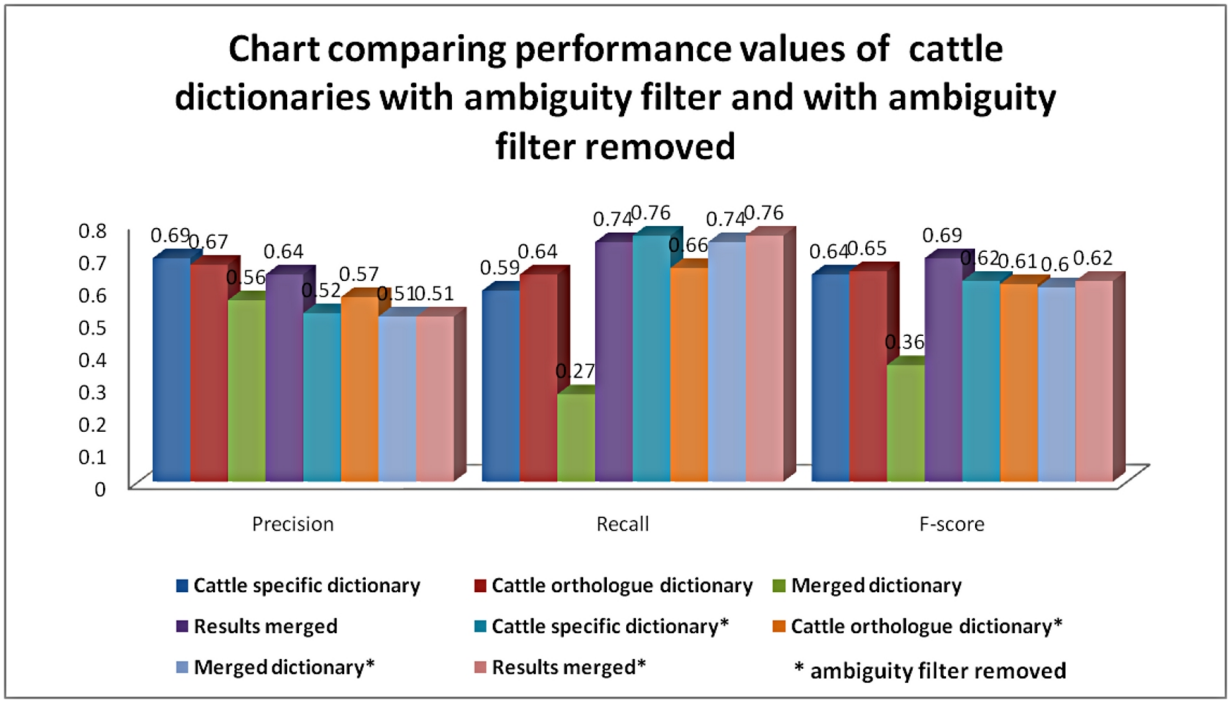


Figure 8.15: Comparison of cattle dictionary performance scores

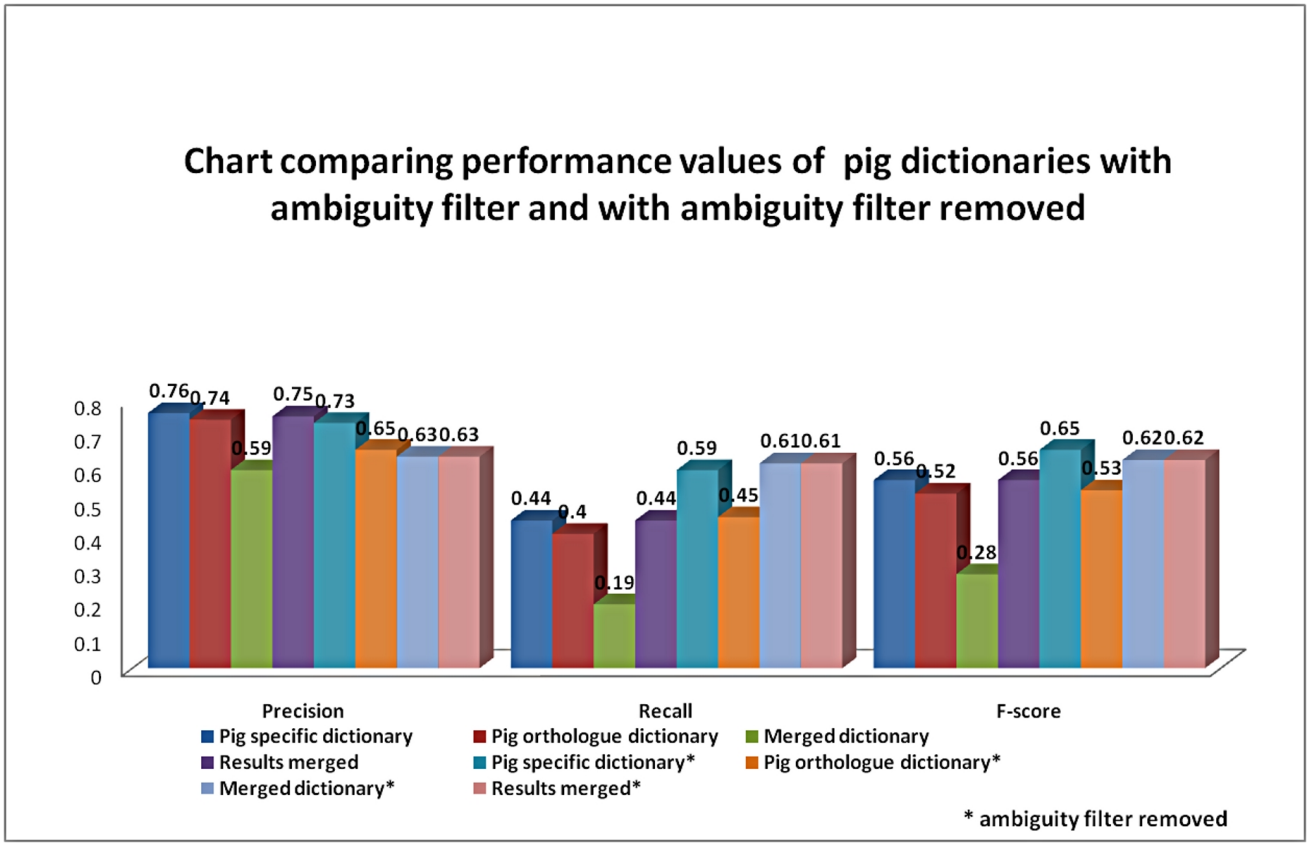


Figure 8.16: Comparison of pig dictionary performance scores

From the graph it can be seen that there is a decrease in the performance score results after the removal of ambiguity filter (see Figure 8.15 and Figure 8.16). This decrease in precision, recall and F-score is as a result of the higher number of false positives identified by the system, despite the increase in the number of true positives identified (see Table 8.9). The increase in the number of true positives identified is also reflected in the increase in the number of spelling variants and case sensitive entities identified by the system with each of the dictionaries. This is especially noticeable in case of merged dictionaries, for which the performance scores and number of entities identified were low with the ambiguity filter and a large increase in the performance scores with the ambiguity filter removed.

8.6.2 Interaction networks

From the networks obtained from SCAIView it can be seen that the SCAIView network proposed more interaction partners than that were found in the experimental data. This can be as a result of text mining and co-occurrence approaches used by the system to obtain networks from texts. This can be illustrated by analyzing interaction partners of cattle preimplantation interaction network suggested by String database using co-

occurrence and text mining techniques. It was found that the database also gave an interaction network similar to SCAIView interaction network (see Figure 8.17), while the confirmed experimental network showed lesser interaction partners (see Figure 8.18). So the additional interaction partners given were considered as false positives.

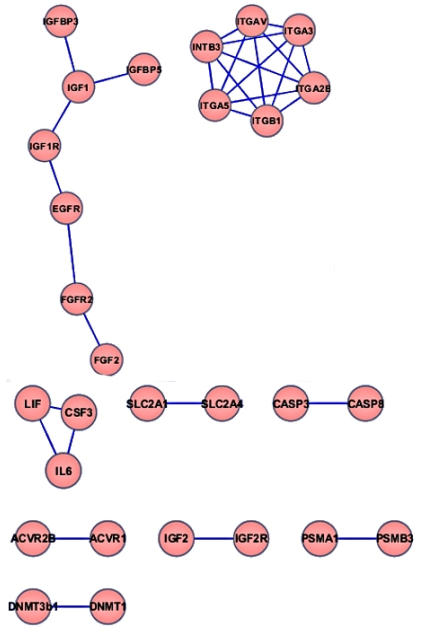


Figure 8.17: Cattle confirmed protein protein interactions

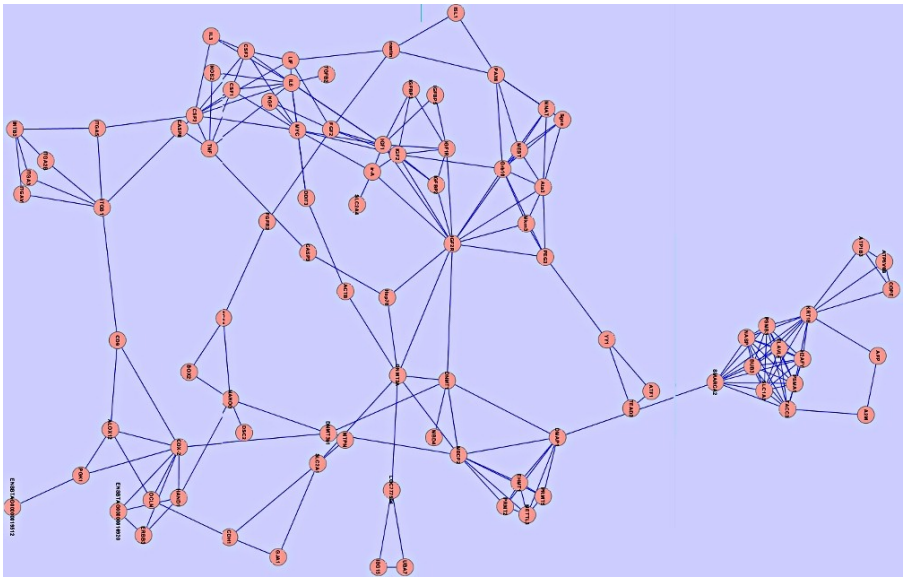


Figure 8.18: Cattle interactions using text mining data from STRING database

The figures show the interactions that are confirmed experimentally and interactions obtained by text mining from String database. The figures show the large variation in the interaction partners found by text mining methods and that are experimentally confirmed. It was found that most of the cattle preimplantation abstracts mention two or more genes in an abstract describing the importance of those in the preimplantation process or the difference in expression of the two genes. So, since the genes were mentioned together, the system considered them as interaction partners following co-occurrence of protein interaction partners.

Experimentally confirmed interaction networks from BIANA contains only about 200 relevant cattle protein data. From protein-protein interaction databases it can be seen that there are far less number of confirmed protein-protein interactions for cattle when compared to humans or mouse. The lesser number of protein-protein interactions for cattle is also reflected in BIANA database, since the database contains interaction data from general protein-protein interaction databases. From the experimental data it can be noted that for cattle proteins, the experimental data from human and mouse genomes are also included since protein interaction network data exclusively from cattle domain is limited.

8.6.3 Preimplantation terminology

When a test search "cattle preimplantaion" was done without using preimplantation terminology and then using preimplantation terminology, it was seen that there was a change in the relative entropy score of gene entities in both of the searches and some gene entities with low ranks in the normal search was ranked high in the search with preimplantation terminology search, due to their difference in the relative entropy score. Since the relative entropy score the found entities in SCAIView are based on the number of entities found and their relative count in the subset corpora selected for the search, for the search with preimplantation terminology, these gene entities were found together with the preimplantation terminology terms and hence had a higher entropy score than the search without preimplantation terminology.

8.6.4 Cattle miRNA targets

Although a list of microRNAs that target a given list of genes can be predicted using mature miRNA sequence, 3' UTR sequences of genes and a prediction algorithm, it cannot be confirmed that the predicted miRNAs are expressed in cattle preimplantation embryo cells. Through prediction a brief list of candidate microRNAs that target a list of genes could be obtained, which can be used as an initial seed points to start an experimental procedure to determine the miRNAs for the gene. During the analysis it was also found that majority of the identified microRNAs and their targets were not found in MicroCosm Targets[4]. The imprecise complementarity of microRNAs to

[4]http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5 last accessed Monday, 26 October 2009

their targets indicate that a single microRNA can target more than one gene, and till now microRNA target prediction systems are based on sequence complementarity and thermodynamics of binding (MIRANDA algorithm) and the reasons for the difference in targets predicted can be due to the difference in miRNA and target sequences selected for analysis, the thermodynamic and statistical factors that were used in algorithm run and the different threshold levels used in cut-off values. The results were agreeing in case of some of microRNAs and their targets and later on it was found that the agreeing results were derived from microRNAs with that had close orthologous groups in mammals (for example, mir-504 with entries found in cattle, human, mouse and rat, let-7b in majority of the mammals).

9 Conclusion

9.1 Summary

This work can be summarized as follows: In the first part researches and research directions in cattle, pig and microRNA genomics were introduced along with the challenges faced. In the follow up second part knowledge discovery and text mining methods were described, mentioning the importance of ontological and semantic search in present day biological text mining. The third part deals with problem definition and text mining needs of animal scientists along with description of background survey on existing databases and bioinformatics applications used in cattle and pig genomics field. The next part deals with the materials and methods, describing the databases and tools used and methodology adopted for generating dictionaries and external database mapping and terminology analysis of the cattle preimplantation period. This part also describes methodology adopted for the mapping of microRNA to their targets. The final part described the results obtained, the performance scores obtained for the dictionaries and some of the analysis done.

The final goals met by this thesis are: generation of cattle, pig and microRNA specific dictionaries and mapping of external database information into SCAIView results for knowledge enrichment, terminology analysis of the cattle preimplantation period, mapping of microRNA target genes to microRNA mappings and integration into SCAIView results, performance analysis of cattle and pig gene dictionaries and analysis of protein interaction networks from SCAIView. This animal science version of SCAIView is intended as a first prototype to demonstrate the possibilities of text mining to the animal scientists, that is entirely dedicated for the used in animal science field.

9.2 Future Prospects

The animal science version of SCAIView will be deployed as a public access version in December 2009 and publications on the basis of the results obtained as a part of this work are planned. In future additions of animal SCAIView, SNP data from experimental results, to mine SNPs and mutation mentions (following the present version of human SCAIView) following data from various micro array databases and ANEXdb database data sets. Another useful addition would be full text version of animal SCAIView, since it was clear that not all gene/protein entities in a full text are mentioned in abstracts and the information content of full text documents are much higher than that of PubMed abstracts, and this is also applicable for microRNAs since it was found that some of the abstracts describes the sequencing project for microRNAs and real microRNA entities are

found only in full text. Relation mining in full texts for microRNAs and possible targets or source gene can also be integrated, since it was noticed that in text were microRNAs are mentioned, normally the microRNA source gene or the targets are mentioned, which could be given as a microRNA target or microRNA source relation, based on various classification algorithms. Similar relationship can be applied to Quantitative Trait Loci (QTLs) in text and marker positions. Apart from named entity recognition techniques, various other approaches in bioinformatics could also be adopted into livestock genomics. One of the candidate fields could be computational systems biology, for simulation of systems such as gene regulatory networks and signal transduction to understand the complex interactions patterns involved in domains such as preimplantation genomics. Since most of the work done in farm animal genomics field is focused on expression analysis of a set of genes, systems biology approaches could be integrated analyze expression levels of different genes in various cellular and culture environments and the resulting data can be integrated into SCAIView, and can be finally used to explore the regulatory network that a candidate gene is involved in, in a particular cell line or during certain stages of growth. A similar approach can be also used for microRNAs by the analysis of microRNA expressions and target gene expressions and integrating the data into gene regulatory networks, depending on the availability of experimental data. Various machine learning approaches could also be adopted into livestock genomics. Using various text mining methods, it could be also possible to extract experimental gene evidenced from texts and augment the present knowledge. By combining such entity -entity relation extraction with experimental data various hypothesis could be formed (based on ABC complementarity), which could further be used as a seed hypothesis for a research direction.

Bibliography

R M Roberts, G.W. Smith, F.W. Bazer, J. Cibelli, G.E. Seidel Jr., D.E. Bauman, L.P. Reynolds, and J.J. Ireland. Farm animal research in crisis. *Science*, 324:468–469, April 2009.

Ulf Leser and Joerg Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings In Bioinformatics*, 6(4):357–369, December 2005.

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics.*, 6: S14, May 2005.

Martin Hofmann-Apitius, Juliane Fluck, Laura Furlong, Oriol Fornes, Corinna Kolearik, Hanser Susanne, Boeker Martin, Stefan Schulz, Ferran Sanz, Roman Klinger, Theo Mevissen, Tobias Gattermayer, Baldo Olivia, and Christoph M Friedrich. Knowledge environments representing molecular entities for the virtual physiological human. *Phil. Trans. R. Soc. A*, 366:3091–3110, June 2008. doi: 10.1098/rsta.2008.0099.

Patricia Laine. *Enhancing Methods for the Biomedical Knowledge Discovery Process.* Master thesis, University of Bonn, August 2008.

Don R Swanson and Neil R Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183–203, 1997.

James E Womack. Advances in livestock genomics: Opening the barn door. *Genome Research*, 15:1699–1705, 2005.

Grahame Bulfield. Farm animal biotechnology. *Trends in Biotechnology*, 18(1):10–13, 2000.

Michael D. Bishop, Stevensen M. Kappes, John W. Keele, Roger T. Stone, Sara L.F. Sunden, Gregory A. Hawkins, Sabina Solinas Toldo, Ruedi Fries, Michael D Grosz, Jakyoung Yoo, and Craig W. Beattie. A genetic linkage map for cattle. *Genetics*, 136: 619–639, 1994.

Josef Fulka,Jr and Helena Fulka. *Somatic cell nuclear transfer.* Somatic Cell Nuclear Transfer (SCNT) in Mammals. Springer New York, December 2007.

Harris A Lewin. The future of cattle genome research: The beef is here. *Cytogenetic and Genome Research*, 102:1015, 2003. doi: 10.1159/000075718.

Ian Wilmut, Lorrain E. Young, and Kevin D Sinclair. Large offspring syndrome in cattle and sheep. *Reviews of Reproduction*, 3:155–163, 1998.

Peter J Hansen and Jeremy Block. Towards an embryocentric world: the current and potential uses of embryo technologies in dairy production. *Reproduction, Fertility and Development*, 16:1–14, 2004.

Christine Wrenzycki, D Herrmann, A Lucas-Hahn, E Lemme, K Korsawe, and H Niemann. Gene expression patterns in in vitro-produced and somatic nuclear transfer-derived preimplantation bovine embryos: relationship to the large offspring syndrome? *Animal Reproduction Science*, 82-83:593–603, July 2004.

A.G. de Vries, A. Sosnicki, J.P. Garnier, and Graham S Plastow. The role of major genes and DNA technology in selection for meat quality in pigs. *Meat Science*, 49:S245–S255, 1998.

David P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116: 281297, 2004.

Yonatan Grad, John Aach, Gabriel D. Hayes, Brenda J. Reinhart, George M. Church, Gary Ruvkun, and John Kim. Computational and experimental identification of c. elegans microRNAs. *Molecular cell*, page Published Online, 2003.

V. Narry Kim. MicroRNA BIOGENESIS: COORDINATED CROPPING AND DICING. *Molecular Cell Biology*, 6:376–385, May 2005.

Tim A. Rand, Krzysztof Ginalski, Nick V. Grishin, and Xiaodong Wang. Biochemical identification of argonaute 2 as the sole protein required for RNA-induced silencing complex activity. *PNAS*, 101(40):14385–14389, 2004.

T. G. McDaneld. MicroRNA: mechanism of gene regulation and application to livestock. *Journal of animal science*, 87(14 Suppliment):E21–E28, 2009.

B.M. Engels and G. Hutvagner. Principles and effects of microRNA-mediated post-transcriptional gene regulation. *Oncogene*, 25:61636169, 2006.

Thimmaiah P. Chendrimada, Kenneth J. Finn, Xinjun Ji, David Baillat, Richard I. Gregory, Stephen A. Liebhaber, Amy E. Pasquinelli, and Ramin Shiekhattar. MicroRNA silencing through RISC recruitment of eIF6. *Nature*, 447:823–828, 2007.

S.P. Chan and F.J. Slack. microRNA-mediated silencing inside p-bodies. *RNA Biology*, 3:97–100, 2006.

Victor Ambros. The functions of animal microRNAs. *Nature*, 431:350–355, 2004.

Christine Esau, Xiaolin Kang, Eigen Peralta, Elaine Hanson, Eric G. Marcusson, Lingamanaidu V. Ravichandran, Yingqing Sun, Seongjoon Koo, Ranjan J. Perera, Ravi Jain, Nicholas M. Dean, Susan M. Freier, C. Frank Bennett, Bridget Lollo, and Richard

Griffey. MicroRNA-143 regulates adipocyte differentiation. *Journal of Biological Chemistry*, 279:52361–52365, 2004.

Jeremy Miles, Tara McDaneld, Ralph Wiedmann, Robert Cushman, Sherrill Echternkamp, Jeffrey Vallet, and Timothy Smith. MicroRNA expression profile in bovine cumulus-oocyte complexes during late oogenesis. *Reproduction, Fertility and Development*, 21 (1):186, 2009.

W. Barendse, A. Reverter, R. J. Bunch, B. E. Harrison, W. Barris, and M. B. Thomas. A validated Whole-Genome association study of efficient food conversion in cattle. *Genetics*, 176(3):18931905, 2007.

Bernd Mueller. *Visualization and Analysis of Extracted Information from Full Text and Patent Corpora*. Master thesis, University of Bonn, 2009.

William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. Knowledge discovery in databases: An overview. *AI Magazine*, 13:57–70, 1992.

Mack Robert and Hehenberger Michael. Text-based knowledge discovery: search and mining of life-sciences documents. *Drug discovery today*, 7:89–98, 2002.

C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, USA, 2nd edition edition, 1979.

Peter Palaga, Long Nguyen, Ulf Leser, and Jorg Hakenberg. High-performance information extraction with AliBaba. volume 360, Saint Petersburg, Russia, March 2009. ACM.

Conrad Plake, Marcus Pankalla, Torsten Schiemann, Jrg Hakenberg, and Ulf Leser. Ali baba PubMed as a graph. *Bioinformatics*, 22(19):2444–2445, April 2006. doi: 10.1093/bioinformatics/btl408.

Dietrich Rebholz-Schuhmann, Harald Kirsch, Miguel Arregui, Sylvain Gaudan, Mark Riethoven, and Peter Stoehr. EBIMedtext crunching to gather facts for proteins from medline. *Bioinformatics*, 23:e237e244, 2006. doi: 10.1093/bioinformatics/btl302.

Robert Hoffmann and Alfonso Valencia. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21(2):ii252–ii258, 2005.

Nicola Guarino. Formal ontology and information systems. In *Proceedings of FIOS'98*, pages 3–15, Trento, June 1998. IOS.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Gavin Sherlock, David P Hill, Laurie Issel-Tarver, Suzanna Lewis, and Gerald M. Rubin. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, May 2000.

Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270, 2004.

Steffen Schulze-Kremer. Ontologies for molecular biology and bioinformatics. *In Silico Biology*, 2:179–193, 2002.

Robert Stevens, Carole A. Goble, and Sean Bechhofer. Ontology-based knowledge representation for bioinformatics. *Briefings In Bioinformatics*, 1(4):398–414, November 2000.

Michael Uschold and Gruninger Gruninger. Ontologies and semantics for seamless connectivity. *ACM SIGMOD Record*, 33(4):58 – 64, 2004.

Russ B Altman, Michael Bada, Xiaoqian J. Chai, Michelle Whirl Carillo, Richard O. Chen, and Neil F. Abernethy. RiboWeb: an ontology-based system for collaborative molecularbiology. *Intelligent Systems and their Applications, IEEE*, 14:68–76, 1999.

W.M. Gelbart, M Crosby, B Matthews, W.P. Rindone, J Chillemi, S. Russo Twombly, D Emmert, Michael Ashburner, R.A. Drysdale, E Whitfield, G.H. Millburn, A de Grey, T Kaufman, K Matthews, D Gilbert, V Strelets, and C Tolstoshev. FlyBase: a drosophila database. the FlyBase consortium. *Nucleic Acids Research*, 25(1):63–66, 1997.

J.M. Cherry, C Adler, C Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y Jia, G Juvik, T Roe, M Schroeder, S Weng, and D Botstein. SGD: saccharomyces genome database. *Nucleic Acids Research*, 26(1):73–79, 1998.

Judith A. Blake, Joel E. Richardson, Carol J. Bult, Jim A. Kadin, and Janan T. Eppig. MGD: the mouse genome database. *Nucleic Acids Research*, 31:193–195, 2003a.

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, the OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Suzanna Lewis, Patricia L Whetzel, Nigam Shah, and Susanna-Assunta Sansone. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25:1251–1255, November 2007.

Christine Golbreich, Matthew Horridge, Ian Horrocks, Boris Motik, and Rob Shearer. OBO and OWL: leveraging semantic web technologies for the life sciences. In *The Semantic Web*, volume 4825/2008 of *Lecture Notes in Computer Science*, pages 169–182. Springer Berlin / Heidelberg, October 2007.

Barry Smith, Werner Ceusters, Bert Klagges, Jacob Khler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46, 2005.

Olivier Bodenreider and Robert Stevens. Bio-ontologies: current trends and future directions. *Briefings In Bioinformatics*, 7:256–274, 2006.

Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44, 2005.

Alexander Pretschner and Susan Gauch. Ontology based personalized search. In *Proceedings. 11th IEEE International Conference on Tools with Artificial Intelligence*, pages 391–398, Chicago, November 1999. IEEE Xplore.

Evelyn B Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, Maslen Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32:D262–D266, 2004.

Andreas Doms and Michael Schroeder. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Research*, 33:W783–W786, 2005.

R Guha, Rob McCool, and Eric Miller. Semantic search. In *International World Wide Web Conference*, pages 700–709, Budapest, 2003. ACM.

Catia Pesquita, Daniel Faria, Andr O. Falco, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLOS Computational Biology*, 5(7), July 2009.

Mingzhu Zhu, Lei Gao, Zheng Guo, Yanhui Li, Dong Wang, Jing Wang, and Chenguang Wang. Globally predicting protein functions based on co-expressed proteinprotein interaction networks and ontology taxonomy similarities. *Gene*, 391(1-2):113–119, 2007.

Zhi-Ping Liu, Ling-Yun Wu, Yong Wang, Luonan Chen, and Xiang-Sun Zhang. Predicting gene ontology functions from protein's regional surface structures. *BMC Bioinformatics*, 8:475, 2007.

Francisco M Couto, Mrio J. Silva, Vivian Lee, Emily Dimmer, Evelyn Camon, Rolf Apweiler, Harald Kirsch, and Dietrich Rebholz-Schuhmann. GOAnnotator: linking protein GO annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, 1:19, 2006.

Francisco M. Couto, Mrio J. Silvia, and Pedro M. Coutinho. Implementation of a functional semantic similarity measure between Gene-Products. November 2003.

David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. GOToolBox: functional analysis of gene datasets based on gene ontology. *Genome Biology*, 5(12):R101, 2004.

Zhidian Du, Lin Li, Chin-Fu Chen, Philip S. Yu, and James Z. Wang. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Research*, published online:15, 2009.

Andy S Law and Alan S Archibald. Farm animal genome databases. *Briefings in Bioinformatics*, 1:151–160, 2000.

Jue Ruan, Yiran Guo, Heng Li, Yafeng Hu, Fei Song, Xin Huang, Karsten Kristiensen, Lars Bolund, and Jun Wang. PigGIS: pig genomic informatics system. *Nucleic Acids Research*, 35:D654–D657, 2007.

Hirohide Uenishi, Tomoko Eguchi, Tetsuya Sawazaki, Daisuke Toki, Hiroki Shinkai, Naohiko Okumura, Hamasima Noriyuki, and Takashi Awata. PEDE (Pig expression data explorer): construction of a database for ESTs derived from porcine full-lenght cDNA libraries. *Nucleic Acids Research*, 32(32):D484–D488, 2004.

Fiona M McCarthy, Nan Wang, G Bryce Magee, Bindu Nanduri, Mark L Lawrence, Evelyn B Camon, Daniel G Barrell, David P Hill, Mary E Dolan, W Paul Williams, Dawn S Luthe, Susan M Bridges, and Shane C Burgess. AgBase: a functional genomics resource for agriculture. *BMC Genomics*, 7:229, 2006.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Marcel Dettling, Ben Bolstad, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Irizarry Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

D Maglott, J Ostell, K.D. Pruitt, and T Tatusova. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33:D54–8, 2005.

Amos Bairoch, Rolf Apweiler, Cathy Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh. The universal protein resource (UniProt). *Nucleic Acids Research*, 33: D154–D159, 2005.

Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

D. P. Wall, H. B. Fraser, and A. E. Hirsh. Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1711, 2003.

Li Li, Jr Christian J. Stoeckert, and David S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13:2178–2189, 2003.

Anton Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora Marks. MicroRNA targets in drosophila. *Genome Biology*, 4:P8, 2003.

Gabriele Varani and William H. McClain. The G.U wobble base pair a fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO reports*, 1(2):1823, 2000.

Stefan Wuchty, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145165, 1999.

Erik Hatcher and Otis Gospodneti. *Lucene in action*. Manning Publications, 2004.

John H. Gennari, Mark A. Musen, Ray W. Fergerson, William E. Grosso, Monica Crubzy, Henrik Eriksson, Natalya F. Noy, and Natalya F. Tu. The evolution of proteegee: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58(1):89–123, 2003.

Philip V. Ogren. Knowtator: a proteegee plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations*, pages 273–275, New York, 2006. Association for Computational Linguistics.

Natalya F. Noy, Michael Sintek, Stefan Decker, Monica Crubzy, Ray W. Fergerson, and Mark A. Musen. Creating semantic web contents with protg-2000. *IEEE Intelligent Systems*, 16(2):60 – 71, 2001.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Idekker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504, 2003.

Juliane Fluck, Daniel Hanisch, Heinz-Theodor Mevissen, and Ralf Zimmer. Playing biology's name game: identifying protein names in scientific text. In *Pacific Symposium on Biocomputing*, volume 8, pages 403–414, 2003.

Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Robert Leaman, Katrin Fundel, Joerg Hakenberg, Chengjie Sun, Heng hui Liu, Rafael Torres, Michael Krauthammer, WilliamW Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K Bretonnel Cohen, and Lynette Hirschman. Overview of BioCreative II gene normalization. *Genome Biology*, 9:S3, 2008.

Thomas Karopka, Juliane Fluck, Heinz-Theodor Mevissen, and Aenne Glass. The autoimmune disease database: a dynamically compiled literature-derived database. *BMC Bioinformatics*, 7:325, 2006.

Juliane Fluck, Heinz Theodor Mevissen, Holger Dach, Marius Oster, and Martin Hofmann-Apitius. ProMiner: recognition of human gene and protein names using regularly updated dictionaries. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, page 149151, Madrid, 2007.

Tobias Gattermayer. *SCAIView Annotation and visualization system for knowledge discovery.* Master thesis, University of Bonn, 2007.

Christoph M Friedrich, Holger Dach, Tobias Gattermayer, Gerhard Engelbrecht, Siegfried Benkner, and Martin Hofmann-Apitius. @neuLink: a service-oriented application for biomedical knowledge discovery. *Studies in health technology and informatics*, 138: 165–72, 2008a.

Ramon Aragues, Daniel Jaeggi, and Baldo Oliva. PIANA: protein interactions and network analysis. *Bioinformatics*, 22(8):10151017, 2006.

Christoph M. Friedrich, Martin Hofmann-Apitius, Robert Dunlop, Ioannis Chronakis, Miriam C.J.M Sturkenboom, Roelof Risselada, Ferran Sanz, and Baldo Oliva. Initial results on knowledge discovery and decision support for intracranial aneurysms. In *Proceedings of HEALTHINF 2008 Conference*, pages 265–272., 2008b.

S Kullback and R.A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

Christine G Elsik, Kim C Worley, The Bovine Sequencing, and Analysis Consortium. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science*, 324:522–527, April 2009.

David S. Roos, Aaron J. Mackey, Christian J. Stoeckert Jr, and Feng Chen. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, 34, 2006.

Judith A. Blake, Joel E. Richardson, Carol J. Bult, Jim A Kadin, and Janan T Eppig. MGD: the mouse genome database. *Nucleic Acids Research*, 31:193–195, 2003b.

Burkhard Rost. Twilight zone of protein sequence alignment. *Protein Engineering*, 12 (2):8594, 1999.

Marius Oster. *Using Latent Semantic Indexing for the Disambiguation of Global Abbreviations in Biomedical Literature.* Master thesis, Fachhochshule Bonn-Rhein-Sieg, 2008.

J.F. Rual, K. Venkatesan, T. Hao, Hirozane-Kishikawa T., A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, N. Simon, N. Boxem, S. Milstein, J. Rosenberg, D.S. Goldberg, L.V. Zhang, S.L. Wong, G. Franklin, S. Li, J.S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R.S. Sikorski, J. Vandenhaute, H.Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M.E. Cusick, D.E. Hill, F.P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, 2005.

YC Lee and YK Lai. Integrity of intermediate filaments is associated with the development of acquired thermotolerance in 9L rat brain tumor cells. *Journal of Cellular Biochemistry*, 57:150–162, 1995.