MilkMine: Text-mining, milk proteins and hypothesis generation

Stephen Edwards BSc. (Hons)

Master of Philosophy

Institute of Structural and Molecular Biology

School of Biological Sciences

University of Edinburgh

2008

Abstract

The vast and increasing volume of biological data can make it a struggle for scientists to keep up-to-date with the latest research and as a consequence they may miss significant biological links, particularly those that extend outwith their own area of expertise. *MilkMine* is an attempt to provide a single informatics resource to help milk protein scientists mine this information mountain more effectively, by integrating standard experimental data types with data generated by emerging text-mining techniques.

A method was initially developed to identify milk-related terminology from peer-reviewed biological literature and this was used to complement the Unified Medical Language System (UMLS), a large thesaurus of biological concepts, their variant names and their types. The resultant enriched ontology was then mapped to the free text of peer-reviewed biological literature using the MMTx program producing a database of semantically enriched sentences.

A co-occurrence relation extraction algorithm was written to identify relationships between milk proteins and peptides, and other biological concepts, such as diseases or biological processes. Using these literature relation sets new hypotheses can be generated using the basic principle that if "*A* is linked to *B*", and if "*B* is linked to *C*" then we can infer an association between *A* and *C*. Filtering and downstream processing of the many generated relationships promotes significant interactions. These literature relations and hypotheses are integrated with biological data into the *MilkMine* database.

The *MilkMine* database is built upon on a generic data warehousing system, InterMine. This tool enabled the integration of traditional data types, such as protein sequence or structural data, from a variety of sources (*e.g.* UniProt). However, the standard InterMine model was also extended by the author to include other data sources (*e.g.* the Protein Data Bank) and to incorporate the output of the text-mining algorithm. This integration of otherwise disparate information allows more complex querying of the data, across many data types. For example, protein sequences are mapped to instances of the names, synonyms or symbols of the protein in text, therefore a raw fragment of amino acid sequence (*e.g.* a particular binding region) can be used to search the *MilkMine* database for literature information as well as the interactions and hypotheses of those proteins that contain the sequence. The *MilkMine* resource is accessible online

(www.bioinformatics.ed.ac.uk/milkmine) through a professional level query interface offering many features such as an interactive query builder, standard ready-to-run queries, bulk downloads and the ability to store user preferences and query histories. Evaluation of *MilkMine* showed that the text-mining algorithm, as well as the data integration, could provide the user with interesting connections for further study.

Declaration statement

I declare that the composition of this thesis is entirely my own work unless otherwise stated in the text. This work has not been submitted for any other degree or professional qualification.

(Stephen Edwards)

Dedication and acknowledgements

I would like to thank my two main supervisors: Professor Lindsay Sawyer and Professor Bonnie Webber for their commitment, patience, time, support and encouragement which have been given freely towards my M.Phil, during the good and the tough times.

A number of people deserve my thanks and praise for their technical support during the work of this thesis. These are, in order of magnitude of contribution:

- The InterMine project team from the University of Cambridge (UK) for support with and on-going development of the InterMine generic database management system, as well as for advice given throughout the implementation of MilkMine. Particular thanks go to Richard Smith and Kim Rutherford who were instrumental in helping me to get MilkMine off the ground. InterMine offered me database and web design expertise far in excess of my own skills and without InterMine, the MilkMine tool would never have been as professional an implementation as it is today. My hope would be that this thesis will be of value to the InterMine team as they progress with their own design and implementation strategy.
- My collaborator, Dr Carl Holt, for his unflinchingly methodical and objective analysis of the MilkMine concept and implementation.
- Dr Alastair Kerr, Shakir Ali and Dr Paul Taylor for technical database and server support and maintenance, including assistance with PostgreSQL, TomCat and Java applications upon which MilkMine relies.
- Professor Sophia Ananiadou for developing and explaining the Termine application (automated terminology extractor algorithm) of which I made use in Chapter 4.

I would also like to thank the many MilkMine evaluators who kindly agreed to beta-test the web interface and provide vital feedback: Matthew Lange (UC Davis); Danielle Lemay (UC Davis); Raphael Flores-Jimenez (Cal Poly); Karsten Qvist (Chr Hansen).

I acknowledge the Biotechnology and Biological Sciences Research Council (BBSRC) for funding this project and the University of Edinburgh for providing the facilities and equipment to allow me to undertake this thesis.

Finally, and most of all, I would like to thank my wife for her loyal support, trust and confidence; without her this thesis would never have seen the light of day. As someone who has all too often been sidelined for the sake of the completion of this work, I am forever in her love and debt.

I dedicate this work to God and trust that it will used for the furtherance of His kingdom.

Overall Thesis Contents

Glossary

ACE	Angiotensin converting enzyme
AF	Atrial fibrillation
ALA	Alpha-lactalbumin
AMP	Anti-microbial peptide
BLG	Beta-lactoglobulin
Corpus	A collection of documents, such as Medline citations (plural form is corpora).
CPP	Caseino-phosphopeptides
CRE	Co-occurrence relation extraction
CUI	Concept unique identifier
FN	False negative
FP	False positive
GO	Gene Ontology
HG	Hypothesis generation
IDF	Inverse document frequency
IE	Information extraction
iHOP	Information Hyperlinked Over Proteins
IMGC	International Milk Genomics Consortium
IR	Information retrieval
LGT	Literature gradient technique
MeSH	Medical Subject Headings
milkER	Milk Extraction Resource
MMSYS	MetamorphoSys program
MMTx	Meta-map transfer program
NCBI	National Centre for Biotechnology Information
NER	Named Entity Recognition
NLP	Natural language processing
P	Precision
PDB	Protein Data Bank
POS	Part of speech
R	Recall
RE	Relation extraction
SIDS	Sudden infant death syndrome
TF	Term frequency
TN	True negative
TP	True positive
UMLS	Unified Medical Language System
Webapp	The web application which provides an interface to the MilkMine database.

Chapter 1 - Introduction

There has been an emerging trend in biological research, particularly over the past two decades, towards experiments which produce large amounts of data, such as yeast 2 hybrid screens, rapid genome sequencing and through use of microarray technology. At the same time there has been a rapid growth in the number of scientific publications deposited in literature databases year on year (see Figure 1.1a). For instance over 750,000 citations[1] were added to the literature database Medline in 2007 alone and which now contains over 18 million citations. Despite the obvious advantages of having increased amounts of data, it raises a new question: How is a laboratory biologist to identify and assimilate all of the data relevant to his field and to use them to their full potential?

1.1 Amalgamation of biological data

In the recent past, creating a database has been a popular method to bring together biological data of a particular type (such as the UniProt database for protein sequence information and the Protein Data Bank (Berman et al., 2000) for 3-D protein crystal structures) or all data related to a niche subject area (such as the Nuclear Protein Database (Dellaire et al., 2003)). One key advantage is that all of the homogenous data is in one searchable space. However, the growing number of biological databases (1078 biological databases are included in the Nucleic Acids Research list of 2008 (Galperin, 2008)) has led to two major problems. Firstly, creating methods able to search, analyse and display the disparate and large volume of information in a user-friendly and intuitive fashion has proved challenging.

The second problem relates to data integration and compatibility. For biological data to become truly meaningful it must be placed within the context of other biological information, for example a raw protein sequence becomes much more meaningful when you know the resulting structure and function of that sequence. Therefore it is essential that it is easy to navigate through different types of biological knowledge so that they can be studied in context. The use of consistent inter-database identifiers has helped this process, however there are still deficiencies of data integration, for example in the mapping of entries in the UniProt protein sequence database to names, synonyms or symbols used in written biological literature.

[1]A search was made on 19th April 2008 through the PubMed interface (www.pubmed.gov) as "all[sb] AND 2007[dp]" search.

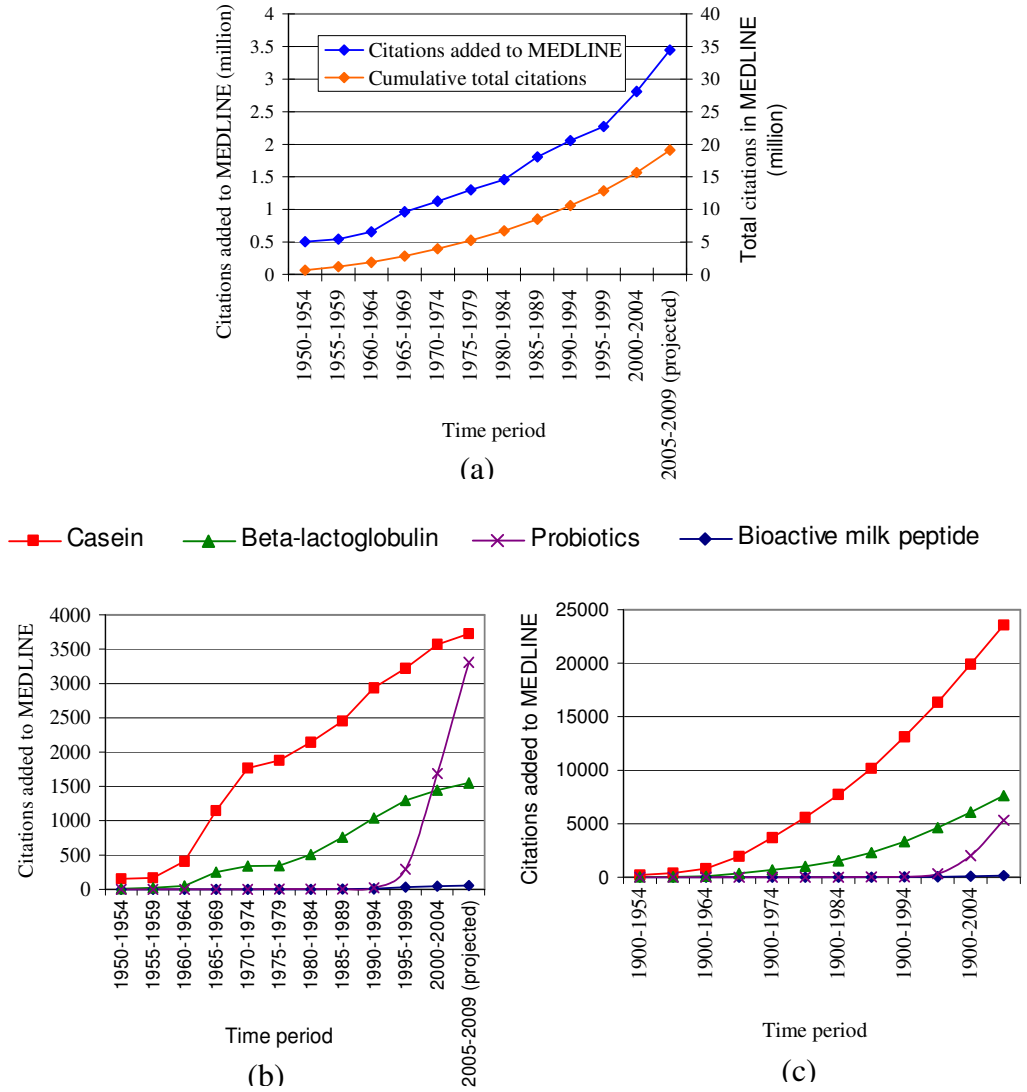


Figure 1.1: (a) Growth in the number of publications added to the literature database, Medline; **(b)** Number of citations added to Medline every 5 years for some key milk-related search terms (Casein, Beta-lactoglobulin, Probiotics, Bioactive milk peptide) and **(c)** Total number of citations.

In light of these two problems a method is needed to assimilate effectively all of the data relevant to a biologist and to present it in such a way that they are not overwhelmed with information. For example, if they wanted to know the key proteins involved in **breast cancer**

and how they are influenced by **nutrition** they may well start with a literature search; however, a search of Medline for the term "breast cancer" retrieves over 100,000 articles with almost 15,000 of these being reviews[1]. A similar search for '("breast cancer") AND nutrition' returns a smaller subset of 2338 articles, although this still includes over 500 review articles[2]. To find the answer to their question, the biologist must start by evaluating the abstracts in the list or by reading through individual publications. The biologist may then want to look at several other databases to identify further information about potential proteins of interest. For example, they may look at UniProt for sequence data, the PDB to view the crystal structure and EnsEMBL to view sequence homologues in other organisms. Thus the biologist has already traversed four different databases through four different interfaces, each of which use different database identifiers. Therefore the amount of time spent sifting through the literature search results and then subsequent databases for information can quickly become prohibitive, particularly if the area studied is outwith the biologist's own area of expertise.

Breast cancer is an extremely well studied area of oncology and therefore it would be expected that a huge number of citations would be returned, however the example still highlights the difficulty of assimilating knowledge from an increasingly large volume of data. Even researching a much more specific example, such as '(invasive breast cancer) AND nutrition' (109 publications) may take a substantial length of time to locate the information required. In a highly specific search, for example 'Phyllodes tumour AND nutrition' no results at all are produced and therefore a more extended period of research would need to be made to examine the link, if any, between the two factors.

In response to these issues of data disparity, overload and integration, database developers have followed two basic approaches: firstly to improve the storage and management of data with greater database compatibility; and secondly to use data-mining algorithms to analyse the data itself more thoroughly. Text-mining, a class of data-mining, uses free text for input and can be applied directly to scientific publications to identify and evaluate information within them. By applying text-mining techniques to biological literature (*i.e.* literature-mining) relationships between biological entities, such as proteins or diseases, can be identified, characterised and compared. Being a computational method text-mining can be performed on large amounts of

[1] Search made on 19th April 2008 through the PubMed interface (www.pubmed.gov).

literature and therefore allows analysis of the identified relationships in a broader context. Unnoticed or undiscovered connections between biological entities can then be made *i.e.* hypotheses can be automatically generated (see Figure 1.2).

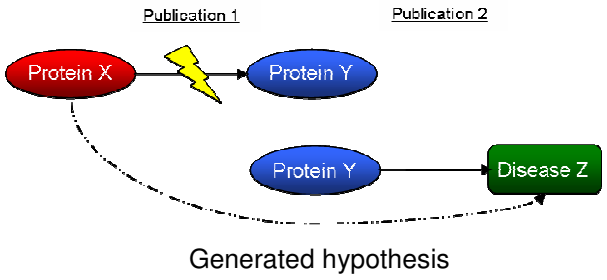


Figure 1.2: Generating hypotheses from biological literature. If Publication 1 states that "Protein X has been shown to activate Protein Y" while Publication 2 states that "Protein Y initiates the pathway of Disease Z"; a hypothesis can be made that Protein X may play a role in the expression of Disease Z.

This technique of generating hypotheses automatically, directly from scientific literature has the potential to increase a biologist's range and capability by making use of information from a much wider area. This in turn will lead to more efficient and perceptive analysis of the growing mass of literature, thereby increasing the potential for scientific discovery. In addition the output can be integrated with other high-throughput experimental data to create a synergy between basic biological data and written knowledge.

One problem with text-mining however, is that there are few impartial evaluation strategies to enable direct comparisons between competing systems (Baumgartner et al., 2008, Bunescu et al., 2005). While a wide variety of text-mining applications cite their performance using the popular metrics **Recall**, **Precision** and **F-score**[1], the tests are often based on different parameters, for example parts of the citation used; citation set sizes or even the initial database search terms or date ranges. These inconsistencies make cross-system evaluation difficult and thus it would be

[1] **Recall** is a measure of how many times an item is identified as a proportion to the total number of instances of that item in the text. **Precision** is a measure of how many times an item is correctly identified as a proportion of all the items identified. **F-score** is a combined value of Recall and Precision = 2 * ((P . R) / (P + R)).

useful to have a structure whereby direct comparisons could be made in an impartial and robust fashion.

1.2 Milk protein science and physiological consequences of diet

Increasingly it is becoming recognised that diet has an extremely important role in health and disease. For example, diet is now known to be related to around 30% of cancers (Beliveau and Gingras, 2007) and has been shown to play a contributory role in inflammatory disease (MacFarlane and Stover, 2007). Milk and milk-derived products constitute a crucial part of our modern diet, being consumed in a variety of forms across all age groups. Many milk proteins and the peptides derived through their digestive processing have been shown to have more than simple nutritional function. These additional functions include hormonal signalling between mother and neonate, antimicrobial strategies within the mammary gland, regulation of gut microflora and remodelling of the mammary gland structure during involution[1] (Clare and Swaisgood, 2000, Hartmann and Meisel, 2007). This is not only true for the major milk proteins (caseins, lactoferrin, β-lactoglobulin, α-lactalbumin, serum albumin and whey acidic protein) but also of the minor milk proteins, implying that these are genuine biological functions (Lonnerdal, 1985). In addition, new components of milk are still being discovered, typified by the recent research into the physiologically active oligosaccharides which show high inter-species variation in composition, complexity and functionality (Boehm and Stahl, 2007).

Milk science has traditionally been undertaken on the macroscale, often looking at the liquid as a whole or at its main constituent parts (*e.g.* milk fat, bulk milk protein). However, this perspective is being replaced by more specific biochemical approaches which make use of informatics and high-throughput based scientific research. This trend complements the recent shift of the multi-billion pound milk and dairy industries from being centred on product towards being an innovative and consumer centric industry. This can be seen particularly in light of the rapid expansion of so-called 'functional foods: foods containing components which have a beneficial impact on target functions in the body beyond those provided by the basic nutrients, minerals and vitamins and which lead to improved state of health and well-being and/or the reduction of risk of disease'. The dairy industry has been a key player in driving this area of research with products claiming the reduction of stress (Spitsberg, 2005), hypertension (Sano et al., 2005,

[1] Remodelling of the mammary gland to its pre-lactational state after lactation has ceased.

Saito, 2008) and cholesterol; and the improvement of digestive health (Ward and German, 2004).

There are, however, potential downsides to increasing the functionality of milk and dairy products whereby the increased physiological activity may inherently present a higher risk to sections of the general population. This note of caution is exemplified by the recent PROPATRIA clinical trial in the Netherlands where a new probiotic yoghurt drink was trialled as a treatment for acute pancreatitis in which 24 out of 152 people (16%) died versus 9 out of 144 people (6%) in the control group (van Santvoort et al., 2006, Besselink et al., 2004, Waterfield, 2008). Therefore, use of *in silico* analysis to draw on and utilise the wealth of current knowledge may become an important addition to biological scientists to create and safely assess new biological hypotheses. Improved use of the data that *is* available will help focus and drive the research programs of the future.

1.3 Project introduction

Due to the expanding, diverse and disparate array of data on milk and milk proteins together with the drive towards an understanding of the physiological function of these dietary components, it was recognized that it would aid researchers to create a single resource database, *MilkMine*. The project aims were to bring together several distinct types of information and to integrate these with the output of text-mining algorithms to allow more complex queries of the data; for example (see Chapter 6):

- Does my milk protein of interest have anti-tumorigenic properties?
- Does my milk protein of interest have any indirect involvement with Sudden Infant Death Syndrome, even though there is no evidence of direct involvement?
- What is the role (if any) of my milk protein of interest in the restructuring of the mammary gland during involution?

Before commencing the project a questionnaire was constructed to identify key information requirements of the milk protein research community. This was completed by researchers having a range of study interests and the results (see Appendix A) were fed into the development cycle as described in Figure 1.3.

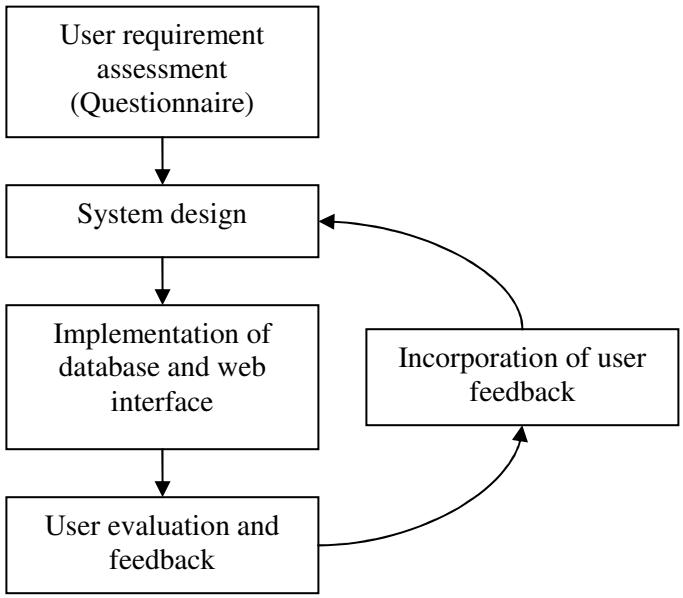


Figure 1.3: Overview of the *MilkMine* database project design and implementation.

This thesis describes the following work:

- ✓ The assessment of MMTx as a tool for Named Entity Recognition (NER, see Chapter 3).
- ✓ The identification of domain related (*i.e.* milk) terminology to enhance the capability of MMTx (see Chapter 4).
- ✓ The creation of a co-occurrence relation extraction and hypothesis generation algorithm which is specifically enhanced for milk protein researchers (see Chapter 5).
- ✓ The amalgamation and integration of disparate biological data with relevant literature-mined data into a single resource with a powerful but easy-to-use web interface (see Chapter 6).

This thesis details the first integration of varied and disparate biological data sets with a scientific literature-mining system that is targeted at a particular sub-domain of biological literature. The author believes that in future the convergence of literature-mining techniques with biological databases will continue to advance and improve. Therefore, it is hoped that this work goes some way to solve the associated problems and can be built upon to further advance the

method in future, either directly through the implementation and expansion of the work covered or by taking into account the discussions and conclusions of this work.

Chapter 2 - Literature review

2.1 Text-mining

The rapid increase in scientific research literature has prompted the development of automated systems to help researchers make use of this information mountain. Text-mining algorithms are a specific category of these systems which extract information from text and store the data in a structured format. Typically they are divided into modules performing different tasks such as part-of-speech taggers (*e.g.* protein [*noun*]), stemmers (*e.g.* activation → *activat*) and entity identifiers (*e.g.* beta-casein [*protein*]); many of these subjects are research fields in their own right. For an in-depth description of text-mining techniques within the biomedical domain please see (Cohen and Hunter, 2004) and (Krallinger et al., 2005).

2.1.1 Information retrieval

The first stage of any text-mining system is information retrieval (IR) where text that matches a specific information request is retrieved from a source (see Figure 2.4). Common examples of IR include searching a literature database such as Medline to return relevant articles (Document Retrieval) or searching a protein sequence database using a specific amino acid sequence.

2.1.2 Named entity recognition (NER)

After information retrieval the next key stage is to identify objects (or entities) within the text; a process known as Named Entity Recognition (NER). Written biological text encompasses several important types of entity such as proteins, genes, RNA, cell types and diseases or disorders. Typically instances of these entities are given names (beta-lactoglobulin) or symbols (BLG) rather than standard database identifiers making it more difficult to associate the information contained within the citation to information in databases (Krallinger and Valencia, 2005). To meet this need a number of systems have been developed to attempt to identify entities automatically, either by using a set of rules to match the names or symbols, statistical methods or by machine learning techniques (see Table 2.1). Rule-based systems usually rely on a combination of regular patterns of text (*e.g.* letter-letter-number) or on large dictionaries of terms. Statistical methods analyse the proportion of terms within documents and between documents to assess their significance. Machine learning-based systems use decision trees and statistical classifiers to 'learn' from current data which terms are entities. Please see (Krallinger and Valencia, 2005) and (Leser and Hakenberg, 2005) for thorough reviews of NER in biological literature.

Table 2.1: Examples of biological Named Entity Recognition (NER) systems.

System	Entities Identified	Type
ABGENE (Tanabe and Wilbur, 2002)	Proteins and genes	Rule-based
NLProt (Mika and Rost, 2004)	Proteins	Rule-based / Machine learning
Termine (Ananiadou, 2006)	Biological terminology	Statistical
ABNER (Settles, 2005)	Proteins, DNA, RNA, cell lines and cell types	Machine learning
GAPSCORE (Chang et al., 2004)	Proteins and genes	Machine learning

2.1.3 Relation extraction

Once entities have been located and characterised in text, relationships *between* them can be identified in a process known as Information or Relation Extraction (RE). Unlike Information Retrieval which simply retrieves relevant documents, RE attempts to drill down into the information that is retrieved (see Figure 2.4). In essence the goal of relation extraction is to extrapolate automatically structured data from unstructured text so that the information can be analysed computationally on a larger scale.

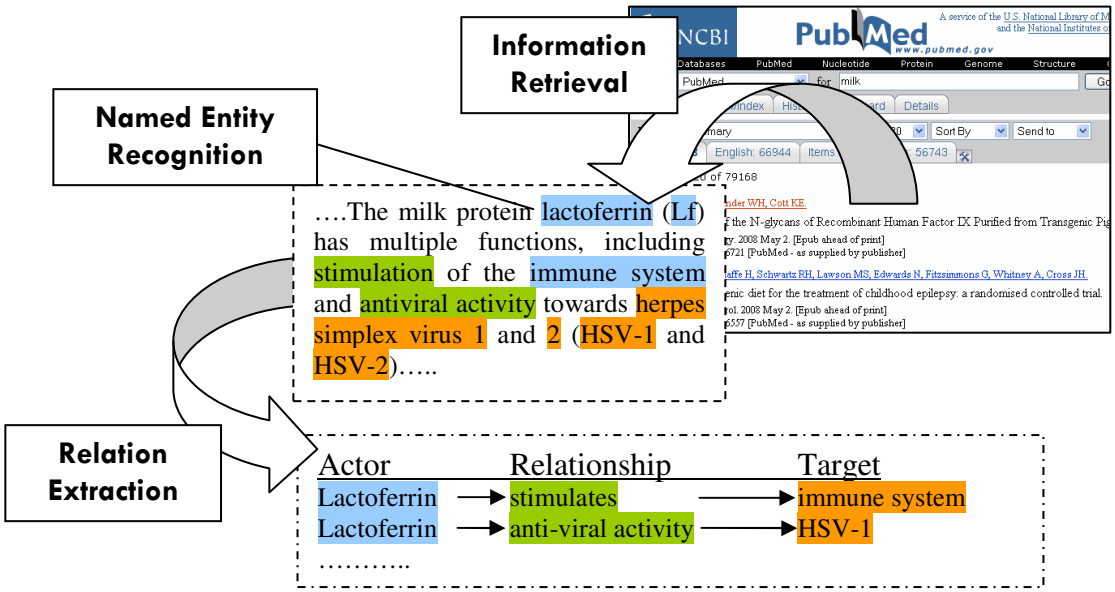


Figure 2.4: Information retrieval (IR) retrieves documents relevant to a query, Name Entity Recognition (NER) locates and characterises entities and then relation extraction (RE) identifies relationships between those entities and records these relationships in a standard format. This information set can then be used as input for further analysis.

Similarly to NER, a number of methodologies have been applied to perform relation extraction which can be placed in three broad categories: co-occurrence based; rule-based or machine learning based systems. Co-occurrence relation extraction makes the assumption that where two entities are located in the same piece of text, there is a relationship between them. Rule-based systems use hard coded patterns to search for instances in text (*[protein]* (0-5 words) *[verb]* (0-5 words) *[protein]* (Blaschke et al., 2002)) while machine-learning systems use statistical methods to 'learn' features of the text which represent relationships between the entities (Bunescu et al., 2005). In reality these methods are not used discretely and many systems use a combination of them, often to perform different parts of the task (Cohen and Hunter, 2008).

2.1.3.1 Co-occurrence based relation extraction

Many studies and systems use the co-occurrence of terms within an abstract, sentence, phrase or clause as indicative of a relationship between those terms (Shah et al., 2003, Hoffmann and Valencia, 2005, Chen and Sharp, 2004, Smalheiser and Swanson, 1998, Pratt and Yetisgen-Yildiz, 2003, Srinivasan and Libbus, 2004, Rebholz-Schuhmann, 2005, Narayanasamy et al.,

2004, Jenssen et al., 2001, Weeber et al., 2003a). PubGene (Jenssen et al., 2001) uses co-occurrence of human gene names within a citation, using the document frequency[1] of the gene pair to assess the strength of the relationship. EBIMed (Rebholz-Schuhmann, 2005) uses co-occurrence within a sentence and ranks these relationships by concept type and the number of supporting sentences. A commonly used co-occurrence based metric is Term Frequency * Inverse Document Frequency (TF*IDF); a statistical method of calculating the importance of a word in a given document against the entire collection of documents. Stephens et al (Stephens et al., 2001) used TF * IDF for single gene occurrences and gene pairs to evaluate the significance of those gene pairs. Narayanasamy et al (Narayanasamy et al., 2004) used a similar technique and produced a visualisation program (TransMiner) to display gene interactions.

Ding et al (Ding et al., 2002) showed that using sentences for co-occurrence relation extraction of biochemical terms from Medline abstracts gave better performance than simply using abstract co-occurrence (see Table 2.2). This result would be expected given that for abstract co-occurrence the average distance between the interacting terms is larger and therefore less likely to represent a true relationship. However, they also found that using sentences gave better performance than using phrases, largely due to the complex nature of natural prose where interacting concepts may appear at opposite ends of a long sentence. Although the use of phrases gave the highest precision the recall was substantially lower[2](Ding et al., 2002). Thus sentences were found to be the optimum unit of text for overall performance of co-occurrence relation extraction, giving sufficient recall of biological relationships while not compromising on precision.

Table 2.2: **Performance of units of text as input for co-occurrence relation extraction (abridged from (Ding et al., 2002)).**

Unit of Text	Recall	Precision	F-Score
Abstract	1.000	0.571	0.727
Sentence	0.849	0.638	**0.729**
Phrase	0.621	0.743	0.677

[1] Document frequency = the number of documents that the search term or terms appear in.
[2]**Recall** is a measure of how many times an item is identified as a proportion to the total number of instances of that item in the text. **Precision** is a measure of how many times an item is correctly identified as a proportion of all the items identified. **F-score** is a combined value of Recall and Precision.

Co-occurrence based systems are relatively easy to implement and have been popular in the production of end-user systems. However there are some drawbacks, for example as the relationship is defined simply as two terms which appear together, there is no knowledge about the directionality of the relationship. For example which protein is activating and which is being activated? This has meant that co-occurrence is not normally used in isolation but is combined with other methods as a part of the overall text-mining system.

2.1.3.2 Rule-based relation extraction

Rule-based relation extraction systems were initially popular as they were relatively straightforward to implement by simply matching regular expressions within text. A small set of simple rules (such as "*[protein]* is phosphorylated by *[protein]*") will capture commonly used phrases and therefore correctly capture a reasonable number of relationships. However, a very high number of rules would be required to catch all of the relevant relationships and the process of manually creating a set of rules to identify them is extremely tedious. Ono and Hishigaki (Ono et al., 2001) developed a rule-based system for extraction of protein interactions from yeast and *E. coli* literature achieving high recall (85%) and precision (94%), however this involved manually constructing rules for every interaction verb and is therefore substantially less transferable. Some systems such as (Corney et al., 2004, Chun et al., 2006) use rule-based algorithms to identify automatically biological relations from text thus reducing the time to build the system and apply it to a different subject area or domain. For example, Huang et al (Huang et al., 2004) align sentences and key interaction verbs as input for a pattern matching algorithm and report reasonable performance figures (recall of 80.0% and precision of 80.5%). Please see (Cohen and Hunter, 2004) for a discussion of rule-based relation extraction methods.

Rule-based relation extraction systems use more in depth analysis of the structure, syntax and semantics of a particular piece of text than co-occurrence based systems. Parsers are now relatively accurate at elucidating sentence structure from text; however the variability and frequent ambiguity of natural language still represents a difficulty, particularly over long and complex sentences. While these features are simple for a human to comprehend, they can cause many problems for computer algorithms; even where the sentence syntax is correctly identified the semantics may not be. Thus the work in biological relation extraction is moving away from rule-based systems towards machine learning approaches to the problem.

2.1.3.3 Machine learning based relation extraction

Machine learning approaches are effectively a progression from semi-automatic rule-based systems where computer programs are 'trained' to recognise relationships between tagged entities in a master text. They are given a set of sentences and are told the features of those sentences (for example the number of words in the sentence and their parts-of-speech) and the algorithm will try to pick out the relationships based upon the information it has been given. The output is then evaluated against the correct answer and the process is repeated until the relation extraction performance is maximized.

Although this is an improvement over writing rules by hand, machine learning algorithms are heavily dependent on the availability of tagged corpora of text for training data. More recently the availability of these corpuses has been increasing in number and quantity[1]; for example the GENIA corpus (Kim et al., 2003) is a collection of 2,000 Medline abstracts which have been manually annotated with biological entities by domain experts. A number of biological text-mining systems have used this technique to perform relation extraction. For example *Chun et al* (Chun et al., 2006) applied machine learning to a corpus of prostate cancer documents and reported a relation extraction precision of 92.1%.

2.1.4 Hypothesis generation

Once entities have been identified in text (Named Entity Recognition, see Section 2.1.2) and the relationships between them have been extracted into a standard format (Relation Extraction, see Section 2.1.3), analysis can be performed over the entire set of relations. One of these analyses is hypothesis generation (or Knowledge Discovery) whose aim is to use the set of relationships gathered by relation extraction to make novel or indirect connections between biological concepts which may not have been previously recorded in the scientific literature (see Figure 1.2).

Hypothesis generation (HG) was first utilized in a groundbreaking study by Don Swanson in 1986 where he noticed that there a complementary overlap between the scientific literature describing the causes and symptoms of Raynaud's Disease, and the physiological effects of fish oil consumption (Swanson, 1986). Swanson was able to show that the symptoms of Raynaud's Disease could be alleviated by the physiological action of fish oil, purely by textual analysis of

[1] http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml

the two literatures. Thus a hypothesis was generated which has been subsequently replicated in other studies (Cole and Bruza, 2005, Gordon and Lindsay, 1996, Weeber et al., 2001, Weeber et al., 2000, Wren, 2004) and has been proven in clinical testing (DiGiacomo et al., 1989).

Hypothesis Generation can be performed in two ways: between two user-defined concepts (Closed Discovery), as Swanson used between Raynaud's Disease and Fish oil, or directly from a single given concept (Open Discovery).

2.1.4.1 Closed discovery hypothesis generation (HG)

Closed Discovery requires an initial proposed hypothesis that there is a link between two concepts; a starting concept (A) and a target concept (C). However, it can then be used to identify a number of novel or unnoticed linking (or intermediate) concepts between A and C. An example hypothesis could be, 'Is there a biological connection between the protein plasmin and mammary gland involution[1]?' A literature search for A is carried out and NER and relation extraction are performed to give a set of relationships for A. This process is repeated for C. The hypothesis can then be examined by looking at literature relationships which have a common intermediate concept (B) between the A-B and C-B literature relation sets identified in the relation extraction process (see 2.1.3). A Closed Discovery HG program will follow the steps below:

1. The user must choose two concepts to form the initial proposed hypothesis (start concept A and target concept C). The HG program will then:
2. Find all significant relationships from the literature between the start concept A and other biological concepts (intermediate concepts B_{A1}, B_{A2}, B_{A3}…B_{An}).
3. Find all significant relationships from the literature between the target concept C and other biological concepts (intermediate concepts B_{C1}, B_{C2}, B_{C3}…B_{Cn}).
4. Identify intermediate concepts that are common between the A-B and C-B relationship sets. By mapping A → B ← C relationships, potentially undiscovered A – C relationships can be identified for further investigation (see Figure 2.5).

[1] Involution is the process of remodelling the mammary gland after lactation has ceased.

2.1.4.2 Open discovery hypothesis generation

Open discovery requires only a single concept of interest as a starting point. Literally, 'generate some hypotheses about [concept A]'. An Open Discovery HG program will follow the steps below:

1. The user must choose a concept to generate hypotheses from (start concept A). The HG program will then:
2. Find all significant relationships from the literature relation set between the start concept A and other biological concepts (intermediate concepts B_1, B_2, $B_3 \ldots B_n$).
3. Find all significant relationships from the literature relation set between each of the intermediate concepts (B_1, B_2, $B_3 \ldots B_n$) identified in step 2 and other biological concepts (target concepts C_1, C_2, $C_3 \ldots C_n$).
4. Identify significant A-B and B-C relationship sets. By mapping A → B and B → C relationships, potentially undiscovered A – C relationships can be identified for further investigation (see Figure 2.5).

While Open Discovery has the advantage of automatically generating hypotheses completely from scratch, it is much more computationally expensive and will potentially produce a huge number of hypotheses. A starting concept will invariably be related to many intermediate concepts, which in turn will be related to many target concepts (see Figure 2.5). Thus, although there is no user input required to complete Steps 2, 3 and 4 of the Open Discovery process, it requires more processing of the set of A → B → C hypotheses that are produced.

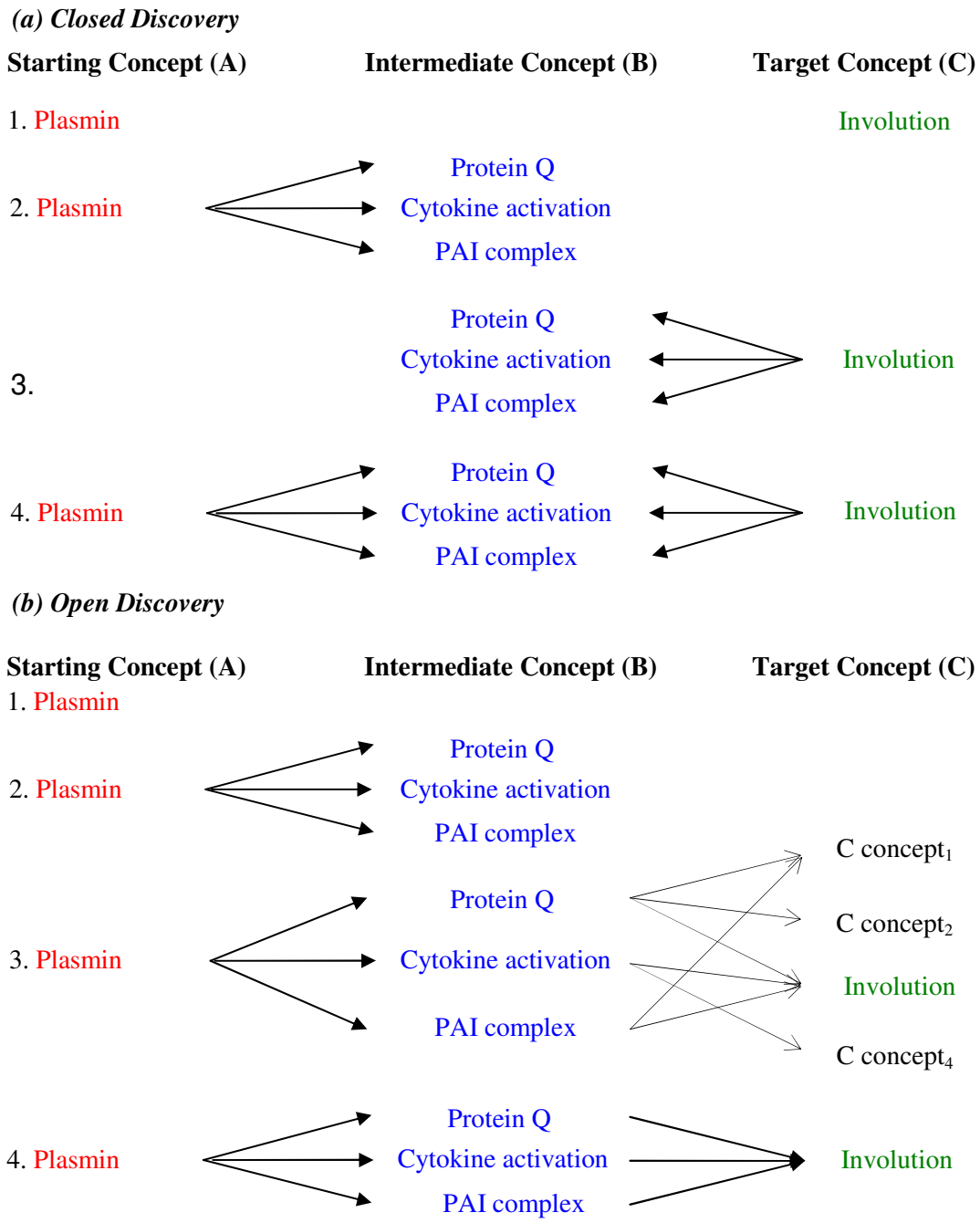


Figure 2.5: **(a)** Closed Discovery. An initial hypothesis can be examined by looking at relationships which have a common intermediate concept between the literature relation sets derived from A literature and C literature. **(b)** Open Discovery. Hypotheses can be generated directly from an initial concept by identifying A → B and then B → C literature relations. By mapping A → B ← C relationships (Closed

Discovery) or A → B → C relationships (Open Discovery), potentially unknown A – C relationships can be identified for subsequent investigation.

A number of biological hypotheses have been automatically generated using Open or Closed Discovery, many of which have been subsequently proven in clinical trails (see Table 2.3).

Table 2.3: Examples of biological hypothesis generation.

Starting Concept	Hypothesis Type	Target Concept	Citations
Magnesium deficiency	Causative	Migraine	(Cole and Bruza, 2005, Swanson, 1988, Weeber et al., 2001)
Long-term endurance training	Causative	Atrial Fibrillation	(Swanson, 2006)
Numerous viruses	Causative	Biological weapon potential	(Swanson et al., 2001)
Turmeric	Therapeutic	Crohn's Disease, spinal cord injuries and retinal diseases	(Srinivasan and Libbus, 2004)
Thalidomide	Therapeutic	Acute pancreatitis, chronic hepatitis C and Helicobacter pylori-induced gastritis	(Weeber et al., 2003a)
Arginine consumption	Therapeutic	Somatomedin C production	(Swanson, 1990)
Indomethacin or oestrogen	Therapeutic	Alzheimer's Disease	(Smalheiser and Swanson, 1996a, Smalheiser and Swanson, 1996b)

2.1.4.3 Intermediate concept restriction

A significant problem common to both Closed and Open Discovery systems is that the set of intermediate concepts produced in Steps 2 and 3 can be very large. Therefore to minimise the amount of manual intervention required, methods to restrict automatically the number of intermediate concepts must be adopted. Many information retrieval or text-mining systems will use a list of common or unimportant words (stop words) to reduce the list of intermediate terms. For Swanson's original and subsequent works the authors manually generated a list of unimportant words (stop words), for example 'the', 'a' or 'where' (Smalheiser and Swanson, 1998), however this can be very time-consuming and domain dependent. Each discovery shown in Table 2.3 was made using different techniques to reduce the intermediate concept (B_n) set.

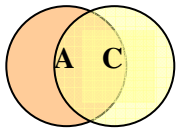
Swanson's (Swanson, 2006) suggestion of a link between atrial fibrillation (AF) and endurance training was made using a method based on Medical Subject Headings (MeSH) and sub-headings. He first looked at population groups with AF using epidemiology subheadings and looked at MeSH terms with "/physiology" subheadings only, giving a much smaller subset to work with. Pratt and Yetisgen-Yildiz (Pratt and Yetisgen-Yildiz, 2003) used a variety of filters to reduce the B concept list such as maximum Medline document frequency; removal of general concepts and restricted the semantic type of the concept allowed in the intermediate list.

Other systems try to avoid the generation of a large number of intermediate terms in the first place. Gordon and Lindsay (Lindsay and Gordon, 1999) used lexical statistics, mainly TF*IDF, with stemming and manual clustering of terms. They then use query expansion using these terms to spider outwards from the start term and were able to recreate Swanson's Raynauds-Fish Oil and Magnesium-Migraine discoveries (Gordon and Lindsay, 1996, Lindsay and Gordon, 1999) however, their technique required a lot of manual intervention. Weeber (Weeber et al., 2003b, Weeber et al., 2001) and Srinivasan's (Srinivasan and Libbus, 2004) methods performed better, both utilizing the semantic knowledge from the UMLS to reduce the B_n concept set. Weeber used Meta-Map (MMTx) on free text to generate hypotheses while Srinivasan used MeSH index terms. The use of a biological vocabulary such as MeSH or the Unified Medical Language System (UMLS) automatically constrains the B_n concept set to biologically relevant concepts only thus removing the need to create or use a stop word list. The downside of using external

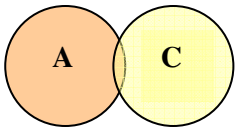
vocabularies is that it places a reliance on external projects which may not be complete or appropriate for a given relation extraction task.

2.1.4.4 Significance of the generated A-C hypothesis

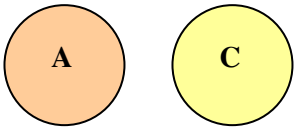
As with Relation Extraction, there needs to be an evaluation of significance of any A-C hypotheses that are produced. For example for a completely novel A-C hypothesis, there should be no articles in the scientific literature describing that A-C relationship. However, often hypothesis generation is used to find further supporting evidence to complement understudied A-C relationships. Indeed, using a weakly known A-C link as the starting point for Closed Discovery makes the task of finding links "several orders of magnitude simpler than an open-ended search" (Swanson et al., 2006). Typically, there will be a gradient of knowledge discovery potential which correlates to the amount of overlap (cohesion) or disparity between the literature describing concept A and the literature describing concept C (see Figure 2.6). Largely overlapping literatures have low discovery potential given that there is a much higher probability that individual researchers will have a reasonable knowledge of both fields and therefore an A-C connection (if significant) may have been published. In contrast, literatures with a degree of small overlap have a high potential for knowledge discovery as the overlap suggests relative closeness of the subject areas while there is only a small amount of current common knowledge between them. Thus there is the potential to find completely new connections between the literatures or to provide further evidence for a small number of potentially understudied connections. Completely disparate literatures have the highest potential for novel hypothesis generation given that there is no crossover between the literatures at all. However, this may be due to the fact that the literatures are in such completely different areas of science (cattle feed production and human cardiovascular anatomy) that no interesting hypotheses could ever be derived from them.



Two literatures with a high degree of cohesion will be unlikely to generate novel biological hypotheses.



A small degree of cohesion between literatures gives a high potential for knowledge discovery.



Disparate literatures have no documents in common and therefore have the highest potential for hypothesis novelty; however the probability of actually producing a hypothesis is lower.

Figure 2.6: Hypothesis generation potential is dependent on the amount of cohesion between the respective literatures (Wren, 2004).

2.1.5 Available text-mining systems

There are a number of end-user text-mining systems available with different functions and features, see Appendix B or (Hunter and Cohen, 2006) for a comparison of end-user relation extraction systems. Commercial companies are also beginning to use text-mining to track movements of competitors and make use of the massive amounts of data generated by their research departments. One example of this is the GCLit system of Astra Zeneca which automatically generates gene summaries using MeSH and gene co-occurrence relation extraction.

2.1.6 Text-mining evaluation

A key stage in developing a text-mining system is evaluation; the author must prove how accurate, relevant and effective their system is. When calculating performance, the following indicators are normally used:

True positives (TP)	instances of a biological entity or a relationship which are correctly identified
False positives (FP)	instances of a biological entity or a relationship which are incorrectly identified

False negatives (FN)	instances of a biological entity or a relationship which are incorrectly ignored

Using these indicators, precision, recall and F-measure of the output of a system can be calculated, to give an absolute value of performance allowing some degree of comparison between competing systems (see Figure 2.7).

Precision (P) = how often the system is correct when it outputs a particular value

$$= \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

Recall (R) = how often correctly finds the right things to output

$$= \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

F-measure = a balance between precision and recall to give a measure of overall system performance

$$= \frac{2 * P*R}{(P + R)}$$

Figure 2.7: Calculation of Recall, Precision and F-measure for the evaluation of Named Entity Recognition (NER) or Relation Extraction (RE) in text-mining systems (Hirschman and Blaschke, 2006).

Often the results will be compared with data that have been manually annotated by human experts, called gold standard data. While the gold standard is often taken to be correct, in reality even human annotators will make mistakes and disagree on all the named entities and relations within the text. Values such as inter-annotator agreement can be used to determine the maximum performance that is possible for a given set of data.

Many Hypothesis Generation papers have used the replication of Swanson's original and subsequent discoveries to evaluate their systems performance (Srinivasan and Libbus, 2004,

Pratt and Yetisgen-Yildiz, 2003, Narayanasamy et al., 2004) however recall and precision figures are not reported in a consistent manner. *Swanson et al* (Swanson et al., 2006) provided a clear definition of precision and recall for Open and Closed hypothesis generation methods, however, this has not been used to compare competing systems.

2.1.7 Abstract vs full-text analysis

Much of the current text-mining and hypothesis generation work has been based on the title, abstract and index (MeSH) terms sections of literature citations. Peer-reviewed scientific abstracts have been shown to contain the greatest density of keywords (Shah et al., 2003) while being relatively small in size and therefore more manageable. Abstracts are also far more widely available and consistently structured than full-text articles.

The full text of an article contains much more data than the abstract alone and the full text of publications are becoming more widely available from literature databases such as PubMed Central and BioMed Central. However, consistency of format and capability of text-mining systems still have some way to go before full-text analysis can become the standard unit of biological literature. Please see (Yeh et al., 2003, Corney et al., 2004, Jose et al., 2007, Sinclair and Webber, 2004, Shah et al., 2003, Kostoff et al., 2004) for work on relation extraction values from different publication sections.

2.2 Integrating data and databases

The growth in biological information has given rise to the problem of interoperability and compatibility between databases. Two key approaches have emerged to meet this challenge: data warehousing and database federation.

Data-warehousing approaches combine disparate data into one database to allow querying of the data through one interface. Examples of end-user biological data-warehouses include EnsEMBL (Hubbard et al., 2002) and FlyBase (Ashburner and Drysdale, 1994). Given the trend towards integrating data in this way, a number of generic data-warehousing software packages have been developed so that researchers can create their own datawarehouse, for example InterMine (Micklem et al., 2006) and Altas (Shah et al., 2005).

Database federation is the process of creating intermediary software *between* databases so that a query can be sent through a single interface to many databases. Examples of database federation clients include Taverna (Oinn et al., 2004) and QIS (Query Integrator System) (Marenco et al., 2004). Although database federation systems have the advantage of not requiring local storage of data and therefore the overheads of maintaining and updating data, they are heavily reliant on the availability and access to the respective sources that they use.

2.3 Terminological resources

A number of projects have looked at creating sets of standard terminology for particular domains to assist the integration of knowledge throughout the scientific community.

2.3.1 Medical subject headings (MeSH)

MeSH[1] is a controlled vocabulary thesaurus, a set of standard biological terms which are assigned to publications in the Medline database by human annotators. This helps to standardise the citations in the database allowing enhanced searching.

2.3.2 Gene ontology (GO)

Three controlled vocabularies in GO, each of which is organised in a directed acyclic graph (DAG): biological processes, cellular components and molecular functions), aim to consistently provide annotation so that independent and distinct databases can include the same semantic interpretation of the annotation.

2.3.3 Unified medical language system (UMLS)

The Unified Medical Language System (UMLS) was founded in 1986 by the National Institute of Health (NIH) to standardise the terms used for biological concepts by combining data from many disparate databases. The key purpose of the UMLS is to link alternative names, terms and lexical variants to a unified concept to improve information retrieval and extraction within the biomedicine and health literature domains. The UMLS is not an attempt to create a single biomedical vocabulary in itself, but to bring together many vocabularies into one standard and useable format.

The UMLS consists of a set of resources and associated software with three main components: a ‘metathesaurus’; a semantic network and lexical interpretation tools.

[1] http://www.nlm.nih.gov/pubs/factsheets/mesh.html

Metathesaurus

The UMLS metathesaurus is a large multi-source and multi-language concept oriented thesaurus and is effectively an amalgam of over 60 databases or vocabularies such as Medical Subject Headings (MeSH), the Gene Ontology (GO), Genbank and the Online Mendelian Inheritance in Man (OMIM) database. For a complete list of all the sources available in the UMLS please see[1]. The wide variety of sources ensures as complete as possible coverage of biological terminology including clinical, public health and experimental biology domains.

Terminology is organized into concepts within the metathesaurus. For example, consider the concept of 'Atrial Fibrillation' which can be defined as the "disorder of cardiac rhythm characterized by rapid, irregular atrial impulses and ineffective atrial contractions"[2]. This condition it is known by a number of different names (or terms) including 'Atrial Fibrillation', 'Auricular Fibrillation' or 'AF', however they all refer to the same concept (see Table 2.4). Thus a concept may be represented by a number of terms - synonyms, abbreviations, spelling variants, case variants or inflectional variants. The UMLS metathesaurus collects all the various terms for a given concept and maps them to a single concept identifier. This makes it an extremely important resource as it can be used by text-mining systems to recognize that 'Atrial Fibrillation' is the same thing as 'Auricular Fibrillation'; a key step in effective text-mining.

[1] http://www.nlm.nih.gov/research/umls/umlsmain.html
[2] Source: MeSH

Table 2.4: Concept, term and string organisation and source transparency in the UMLS metathesaurus.

UMLS concept	Alternative terms for the concept	Alternative names (strings) for the term	Alternative sources of the strings
Atrial Fibrillation	Atrial Fibrillation	Atrial Fibrillation (preferred)	Atrial Fibrillation (from MeSH)
			Atrial Fibrillation (from the Thesaurus of Psychological Index Terms)
		Atrial Fibrillations	Atrial Fibrillations (from MeSH)
	Auricular Fibrillation	Auricular Fibrillation (preferred)	Auricular Fibrillation (from the Thesaurus of Psychological Index Terms)
		Auricular Fibrillations (plural variant)	Auricular Fibrillations (from MeSH)

Where available, the following information is also gathered from the source vocabulary: name and preferred name; synonyms; hierarchy; definition and identifier. Every concept is given a unique 8-digit identifier (*e.g.* C0004238) although source transparency is always maintained by using additional identifiers to keep a record of which source vocabulary the term has come from. This ensures that each term can be traced back to its original source vocabulary (see Table 2.4).

Semantic network

The UMLS Semantic network is a hierarchical set of 135 categories of UMLS concept, called semantic types (see Appendix C). It also defines 54 relationships between those types. Each concept within the metathesaurus is categorised into at least one of 135 semantic types, using the most specific semantic type available. For example the concept 'Diabetes mellitus' is assigned the semantic type [Disease or syndrome].

Lexical tools

The UMLS lexicon is a database of morphological, lexical and orthographic information for the common English language and biomedical vocabulary. Lexical programs are also available, for example to detect and differentiate case; word order and inflectional variation in natural language are used to generate indexes on the metathesaurus. Although these tools are designed to work with the UMLS dataset, they can be used independently on any appropriately formatted data.

Additional Tools

Given its wide scope, the complete UMLS is cumbersome and therefore it must be filtered (customised) to be made applicable for a particular purpose, using the MetamorphoSys (MMSYS) program. This tool also allows the storage of user preferences in a configuration file so that updating the dataset to future UMLS releases can be directly run from this file, thus making updates much more viable.

There is also an online service, the UMLS Knowledge Server (UMLSKS) which is available through an Application Program Interface (API) or a web browser. Alternatively the UMLS can also be loaded into a relational database (Oracle and MySQL are officially supported).

2.3.4 MetaMap transfer (MMTx)

The MetaMap Transfer (MMTx) project was initiated in the early 1990's to create a tool to identify UMLS concepts within free text, thus allowing the integration of the UMLS into text-mining software. MMTx (Aronson, 2001) splits documents into sentences, phrases, terms and words (or tokens) using the MedPostSKR (Smith et al., 2004) part-of-speech tagger (see Figure 2.8). Variant forms such as synonyms, spelling variants, inflections, acronym and abbreviations of the tokens in each phrase are then retrieved from the UMLS lexicon. Each variant form is given a score according to the number of transformations which were required to get from the original form to the variant form.

The program then matches the phrases to the most similar or best covering UMLS metathesaurus strings in the lexicon. If the phrase is not completely covered by one candidate UMLS string, combinations of candidates are put together to best cover the phrase. These candidates are then

evaluated against each phrase based on several criteria (please see (Aronson, 2001) for a thorough description of the MMTx program):

1. Centrality of the term within the noun phrase (for example a term at the head of the noun phrase will carry more weight).
2. The degree of variation between the term as compared with the UMLS metathesaurus string.
3. How many of the noun phrase tokens are covered in the candidate concept, and
4. The cohesiveness of the term compared to the UMLS string (*i.e.* a measure of how well the word order of the term matches the UMLS string).

Candidates and combinations of candidate terms are combined and an evaluation of the mapping is made. An exact match receives a score of 1000; 0 equates to no match.

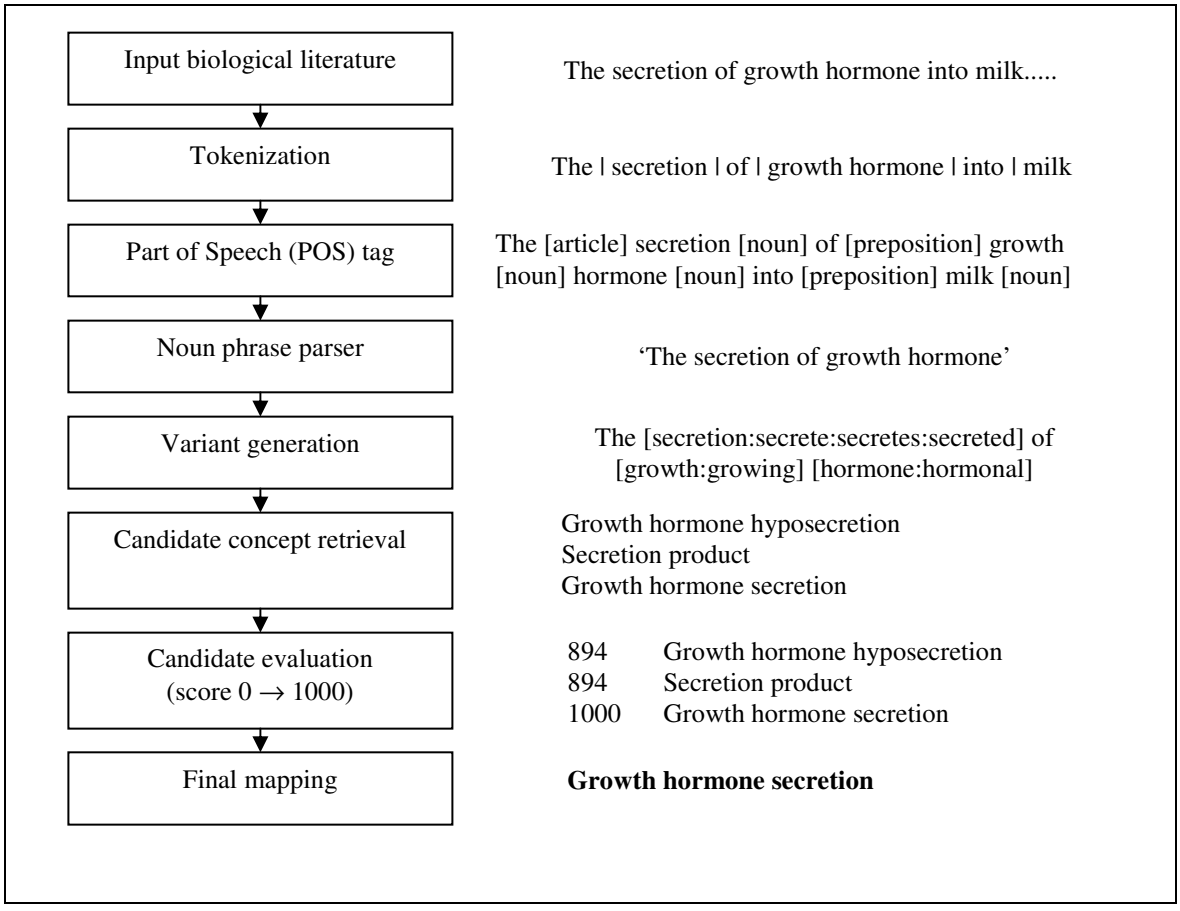


Figure 2.8: The workflow of the MMTx program which matches (or tags) UMLS concepts to free scientific text.

Although the MMTx program was designed to be used with the UMLS, it can be applied to any dataset of concepts provided that the data is in the UMLS format. There are also a number of options available to control and tune the performance of MMTx to suit the purpose of individual users; for example you can limit the candidate concepts to specific semantic types only (only identify gene names within text).

2.4 Biological data resources

There are a number of well established resources of biological data available (see Table 2.5).

Table 2.5: Biological data resources.

Data source	Description
UniProt (2008)	The UniProt database (Universal Protein Resource) is an amalgamation of four source databases of protein sequence data with annotation (function, classification and cross-references).
InterPro (Apweiler et al., 2000)	InterPro is an integrated database of computer annotated information for protein families, domains and sites databases, compiled by the EBI from many smaller sources databases such as Pfam and PROSITE. These source databases use different techniques and use a varying amount of biological information on proteins to derive protein annotation.
Protein Data Bank (PDB) (Berman et al., 2000)	The Protein Data Bank is the largest database of experimental 3-Dimensional data for the structures of proteins and nucleic acids.
IntAct (Hermjakob et al., 2004)	The IntAct database is a collection of thousands of known molecular interactions which have been gathered from scientific publications or from direct user submission.
EnsEMBL (Hubbard et al., 2002)	The EnsEMBL project aims to provide computational annotation of many eukaryotic genomes, including those of human, cow, mouse and rat.

2.5 Milk literature review

2.5.1 Introduction

Milk represents a vital boundary in life between birth and self-sufficiency and is produced by all 4,500 mammalian species. Milk is composed of varying levels of fat, carbohydrate, proteins, minerals, all essential vitamins, leukocytes, growth factors, enzymes, enzyme inhibitors, immunoglobulins, and anti-bacterial proteins and peptides. Although milk is produced primarily for infant nutrition, milk components also perform important physiological functions such as immunoregulation, growth and hormonal responses and antibacterial activity. Therefore they can influence the health of neonate and mother as well as maternal milk production (Farrell and Thompson, 1990, Meisel, 1997).

The composition of milk is highly dependent on the specific needs of the particular species, providing ideal nutrition for the infant. For example, human milk has a protein content of 0.9g/100ml protein (growth rate of 120-180 days to double birth weight), compared to 8.1g/100ml protein in rat milk (growth rate of 2 days to double birth weight) (Hambraeus and Lonnerdal, 2003). Aardvark milk and the milk of cetaceans (whale and dolphin family) have high concentrations of fat due to the scarcity and salinity of water in their respective habitats (White et al., 1985).

2.5.2 Lactation

Lactation (or lactogenesis) is hormonally controlled. During pregnancy the mammary gland differentiates in preparation to secrete milk, however, production is held in check by high progesterone and oestrogen levels (Buhimschi, 2004). When the placenta, the source of progesterone during pregnancy, is removed after child birth copious milk production begins. This is secreted from the alveoli or acini cells within the mammary tissue and then passes to the lactiferous sinuses for ejection (Buhimschi, 2004).

Milk composition typically changes after the initial few days as the nutritional need of the infant changes. Colostrum (early milk) contains much higher levels of immunoglobulins and protein, but less fat than mature milk. This temporal expression is particularly pronounced in marsupials where the milk composition changes over the entire course of suckling as the infant is much less developed at birth than other mammals (Demmer et al., 2001).

2.5.3 Milk proteins

Milk proteins are a key component of milk and include a variety of functional groups including growth factors, hormones, enzymes and enzyme inhibitors. The major milk proteins (α_{s1}-, α_{s2}-, β- and κ- caseins, β-lactoglobulin and α-lactalbumin) are well characterised for several species as small, stable proteins, although the flexible structure of caseins makes them more susceptible to proteolytic cleavage than the highly resistant β-lactoglobulin (Jayat et al., 2004). For a thorough description of the structure and function of the major milk proteins, please see the following review (Swaisgood, 2003).

2.5.3.1 β-lactoglobulin (BLG)

β-lactoglobulin (BLG) is the major milk whey (*i.e.* non-casein) protein of many species such as ruminants and baboons (Perez and Calvo, 1995). However, it is notably absent in human, rodent and lagomorph milks (Perez and Calvo, 1995, Kontopidis et al., 2004, Azuma and Yamauchi, 1991). BLG belongs to the lipocalin family (Ganfornina et al., 2000) which are typically transport proteins, and are also expressed in saliva (as biosensor proteins) and tears (Lacazette et al., 2000). BLG binds many hydrophobic ligands such as retinol (vitamin A), in a central cavity and thus has been hypothesised to be a transport/uptake protein for nutrients (Perez and Calvo, 1995, Perez et al., 1989). Other suggested functions include enzyme regulation (Perez et al., 1992) and neonatal passive immunity although these functions are not conserved across species, however, the genuine function of BLG remains undetermined (Kontopidis et al., 2004, Kontopidis et al., 2002, Flower et al., 2000). It may be that BLG is primarily as an important source of amino acids for the neonate and that this function arose from gene duplication of glycodelin, in which case the other functions are merely a fortuitous by-product (Kontopidis et al., 2004). However as β-lactoglobulin is very resistant to proteolysis and species distribution it seems unlikely that the protein has a strictly nutritional role (Stapelfeldt et al., 1996).

2.5.3.2 Alpha-lactalbumin

α-lactalbumin (ALA) is a small, well characterised metallo-protein with a primary and tertiary structure similar to lysozyme. ALA catalyses the first step of the biosynthetic pathway of lactose, the milk concentration of which is directly related to α-lactalbumin concentration (Brew, 1969). As lactose creates ~50% of the osmotic pressure of milk, the regulation of α-lactalbumin must be tightly controlled (Brew, 1969). Although α-lactalbumin is produced in mammary gland very small amounts leak into maternal bloodstream, this increases during pregnancy and can be used as an indicator of mammary gland development (McFadden et al., 1987). It has been noted that α-lactalbumin is found at very low levels (or is non-existent) in marine mammals (Reich and Arnould, 2007).

2.5.3.3 Caseins

The caseins are a family of natively unfolded proteins and are present in all mammalian milk forming macro structures called micelles (Swaisgood, 2003). These micelles trap calcium, allow the super saturation of calcium found in milk and are vital to the control of pathological calcification in the mammary gland (Swaisgood, 2003, Hambraeus and Lonnerdal, 2003). Casein

micelles vary is size between species through a great diversity of genetic alleles, however, the genomic organisation of the casein gene locus appears to be conserved throughout species (Martin et al., 2003).

2.5.3.4 Minor proteins

Many other proteins are found in milk at low concentrations, for example immunoglobulins. Immunoglobulins in milk are primarily responsible for the passive immunity of the neonate, particularly in colostrums where immunoglobulins constitute 70 - 75% of total milk protein, (Fox, 2003a). Milk also contains many antimicrobial peptides and proteins, such as lactoferrin which binds iron thus starving the bacteria of this vital element.

It has been shown that the total antimicrobial value of milk is greater than the sum of the activity of the immunoglobulins and other host defence proteins which is probably due to synergistic actions and enzymatically released peptides (Clare and Swaisgood, 2000). In this way the milk proteins provide species specific protection against bacterial and viral infection (Florisa et al., 2003) while promoting the establishment of a protective gut microflora (Fanaro et al., 2003, Oddy, 2002, Hanson and Korotkova, 2002, Wold and Adlerberth, 2000). The mammary gland also protects itself from microbial infection through the action of milk proteins, immunoglobulins, iron binding proteins (especially lactoferrin) antimicrobial peptides and other minor proteins (Fox, 2003a).

2.5.4 Bioactive milk peptides

Peptides enzymatically released from milk proteins have also been shown to have biological activity and are known as ‘bioactive peptides’ (see Table 2.6). For example, antiviral activity of peptides derived from milk proteins have been shown against HIV and human cytomegalovirus (HCMV), while the parent milk proteins have no effect (Florisa et al., 2003). Several bioactive peptides can be released from lactoferrin which have been shown to be active against both gram positive and gram negative bacteria *in vitro* (Clare and Swaisgood, 2000). For example, pepsin digestion of lactoferrin releases a potent antibacterial peptide, lactoferricin B, which was shown to be active at a concentration significantly below that of the lactoferrin hydrolysate or lactoferrin itself. These properties appear to be associated with the net positive charge of Lactoferricin B which may kill susceptible micro-organisms by increasing cell permeability *i.e.* with a similar action to anti-microbial peptides (AMPs) (Clare and Swaisgood, 2000). Please see

Meisel (Meisel, 1997) and Clare and Swaisgood (Clare and Swaisgood, 2000) for thorough reviews on bioactive milk peptides.

Table 2.6: Functional classes of bioactive peptides released from milk proteins by enzymatic hydrolysis. Table abbreviated from (Clare and Swaisgood, 2000).

Physiological function	Example protein	Milk protein fragment	Proven activity	Release agent (enzyme)
Antimicrobial activity	Lactoferricin B	Lactoferrin (17-41)	Gram +ve (including MRSA) and –ve bacteria, yeast	Pepsin
Antifungal activity	Isracidin	α_{s1}-casein (1-23)	Candida sp.	Chymosin, chymotrypsin
Antihypertensive activity	α_{s1}-casokinin-5	α_{s1}-casein (23-27)	ACE inhibitor	Proline endopeptidase
Antithrombotic activity	Casoplatelin	κ-casein (106-116)	antithrombotic	Trypsin
Caseino-phosphopeptides	Caseino-phosphopeptide	α_{s1}-casein (59-79)	Calcium binding and transport	Trypsin
Immunomodulation	Lactoferricin B	Lactoferrin (17-41)	Immuno-modulation (+)	Pepsin
Opioid (agonist)	β-lactorphin	β-lactoglobulin (102-105)	Opioid agonist	Trypsin
Opioid (antagonist)	Casoxin C	κ-casein (25-34)	Opioid antagonist	Trypsin

2.5.4.1 Antihypertensive peptides

In its active state the serum protein angiotensin converting enzyme (ACE), induces a rise in blood pressure and has therefore been linked with heart disease. ACE inhibitor sequences have been identified in human caseins and ACE inhibitors have been found in tryptic digests of bovine milk. For example, the tryptic digestion of β-lactoglobulin releases lactorphin peptides with ACE inhibition activity (Clare and Swaisgood, 2000). However, the study suggests that the *in vivo* anti-hypertensive activity of the milk derived peptides IPP, VPP and KVLPVP is *not*

through ACE inhibition directly as they only weakly inhibit ACE. Therefore they may have an altogether different mechanism of action than was previously thought (Fuglsang et al., 2003).

2.5.4.2 Antithrombotic peptides

On the molecular level, the clotting of blood and milk show remarkable similarities (Lemkin et al., 2000). Bovine κ-casein and human fibrinogen display remarkable similarity while trypsin hydrolysis of bovine κ-casein has been shown to inhibit fibrinogen binding to the platelet surface, thus resulting in antithrombotic activity (Clare and Swaisgood, 2000). Patients with heart failure are at higher risk of having problems with thromboses (Myocardial infarction) and therefore may benefit from the properties offered by the bovine κ-casein digest. Many of these patients also have underlying ischemic heart disease and are therefore treated with aspirin, however, aspirin may interact with ACE inhibitors and may block prostaglandin production by the kidney, both resulting in increased blood pressure and therefore risk of heart attack (Verheugt, 2004). The antithrombotic properties of milk peptides may provide a safer, healthy alternative treatment to drugs in patients with heart failure. Although these properties of milk proteins have been known for many years, a thorough trial in this context has not been suggested or undertaken.

2.5.4.3 Casein phosphopeptides (CPP)

Casein phosphopeptides (CPP) are released by trypsin digestion of α_{s1}- α_{s2}- and β-casein and may function as transport molecules for different minerals. CPPs can form complexes with calcium phosphate in the intestine, increasing the absorption of calcium across the small intestine (Clare and Swaisgood, 2000).

2.5.4.4 Opioid peptides

Opioid peptides can be released from milk proteins by enzymatic digestion. A tyrosine residue at the amino terminal end with another aromatic amino acid produces a structural motif in these peptides that allow the peptide to bind to opioid receptor (Saito, 2008). This interaction can result in increased gastrointestinal transit time, modulation of the intestinal transport of amino-acids and stimulation of insulin and somatostatin secretion.

2.5.4.5 Physiological activity

Even nutritionally insignificant amounts of bioactive milk peptides may be sufficient to exert physiological effects (Maeno et al., 1996). Evidence has shown that although they are more

likely to act on local sites in the gastrointestinal tract they can also be absorbed intact and therefore have an increased bioactivity potential within the body (Kayser and Meisel, 1996). Bioactive milk peptides have even been shown to be pass through the blood-brain barrier, interacting with receptors in the brainstem and influencing brain activity (Sun et al., 2003).

Certain peptides from milk can also directly influence the mother, for example, casomorphins can be enzymatically released in the mammary gland and may participate in the endocrine regulation of pregnancy (Koch et al., 1988).

However, care should be taken over results obtained *in vitro* as they may represent a very different situation to that *in vivo*. Even if the reactions are similar, the biological activity of the bioactive peptides must be retained until it reaches the target site, and in a high enough concentration to elicit a physiological effect. For example, the activity of lactoferricin B has been shown to be reduced *in vivo* by the addition of whole milk or mucin (Clare and Swaisgood, 2000). Also, posttranslational modifications can greatly affect the activity of bioactive peptides activity *in vivo*. Therefore bioactive function cannot necessarily be directly attributed to the amino acid sequence (Dziuba et al., 1999).

2.5.4.6 Economic value

There has been a recent surge of interest in bioactive milk peptides from the dairy and healthcare industries (Clare and Swaisgood, 2000). They are now being produced commercially and used as nutraceutical food additives. For example, tryptic hydrolysates of casein containing ACE-inhibitory peptides have been added to yoghurts and milk products which have been shown to reduce hypertension (Fuglsang et al., 2003). These natural additives may be more generally acceptable than artificial drugs in the treatment of diarrhoea (casomorphins), hypertension (casokinins), thrombosis (casoplatelins) dental and bone disease as well as mineral malabsorption (casein phosphopeptides) and immunodeficiency (immunopeptides) (Meisel, 1997).

2.5.5 Milk proteins and disease associations

Composition of the milks of different species varies widely and there can be problems associated with cross-species consumption. For example, in breastmilk the casein content is largely β-casein, whereas in bovine milk this is largely β and α_{s1}-casein. Bovine milk also contains a large

concentration of β-lactoglobulin while breastmilk contains none. These differences are particularly important for the nutrition of infants, for example the infant brain requires arachidonic acid and docosahexaenoic acid (AA and DHA); both of which are found in human milk but are absent in cows milk (please see (Gartner et al., 2005) for a review of the issues surrounding breastfeeding as summarised in Table 2.7.).

However, longer-term problems can also occur. A notable but highly disputed study by Elliot et al (Elliot et al., 1997) suggested a link between a bioactive peptide released from a particular genetic variant of bovine β-casein and the development of Type I diabetes in infants (Elliot et al., 1997). The A^1 and B variants of bovine β-casein have a histidine at position 67, allowing the enzymatic release of a seven amino acid opioid peptide (β-casomorphin7) which has been shown to inhibit human intestinal lymphocyte proliferation (*in vitro*) (Schrezenmeir and Jagla, 2000). It has been suggested that the resultant immune suppression may influence the development of gut-associated immune tolerance, or reduce defence mechanisms towards enteroviruses, both of which have been implicated in Type I diabetes (Lawlor et al., 2005, Elliott et al., 1999). The A^2 bovine β-casein variant does not release β-casomorphin7 under enzymatic digest (Elliot et al., 1997). More recently, high intakes of milk protein rather than meat protein, were also found to increase insulin resistance in 8-year old males (Hoppe et al., 2004a, Hoppe et al., 2004b, Hoppe et al., 2005). Please see Truswell (Truswell, 2005) for a critical review of the alleged bovine β-casein /diabetes link.

Table 2.7: Benefits of breast-feeding (Abridged from (Gartner et al., 2005)).

Beneficiary	Benefit
Infant	Decreases incidence and/or severity of a wide range of infectious diseases, including RTI, UTI, bacterial meningitis
	(Possible) decreased risk of sudden infant death syndrome
	Decreased risk of type I and II diabetes, lymphoma, leukaemia, Hodgkin disease, obesity, hypercholesterolemia and asthma in older children and adults who were breastfed
	Enhanced cognitive development (for certain genotypes)
Mother	Decreased postpartum bleeding
	More rapid uterine involution (due to increased levels of oxytocin)
	Decreased menstrual loss
	Earlier return to prepregnancy weight
	Decreased risk of breast and ovarian cancer
	(Possible) decreased risk of postmenopausal osteoporosis

Consumption of bovine milk has been linked to a number of other problems, such as allergic T-cell reactions (Kaila et al., 1994, Selo et al., 1999, Wroblewska et al., 2004), an increased incidence of asthma, as well as respiratory tract and ear infections, in infants fed with substitute milk (Oddy, 2004). It has also been proposed that bovine hormones in milk (particularly IGF-1, a hormone identical to the human equivalent) may remain active after consumption (Danby, 2005).

2.5.6 Milk genomics

The bovine genome (*Bos taurus*) was sequenced in 2004 (Band et al., 2000, Lewin, 2003) and as the first agricultural animal to be sequenced is important for the understanding of lactational processes.

The genotypic variants of the major milk proteins (caseins and β-lactoglobulin) have been well characterised (Martin et al., 2002) and these alleles may also affect the expression of other abundant milk proteins (Lum et al., 1997, Folch et al., 1996, Elsen, 2005, Bobe et al., 1999, Barillet et al., 2005). Proteomics studies, largely using electrophoresis and mass spectroscopy techniques, have been performed on bovine, murine, caprine, porcine, wallaby and human milks (Kim and Jimenez-Flores, 1994, Molloy et al., 1997, Murakami et al., 1998, Yamada et al., 2002, Charlwood et al., 2002).

2.5.7 Milk informatics prior work

There is a number of milk and milk protein science informatics studies in the current literature. The International Milk Genomics Consortium (IMGC) Portal[1] (German et al., 2006) project was started in 2004 to unite public and private data on experimental milk genomics. This project has a similar goal to *MilkMine* however there is no text-mining component. The Lipgene project[2] is an EU Sixth Framework initiative looking at the interaction of diet and metabolic syndrome. The project runs from 2004 – 2009 and is analysing existing dietary, biochemical, clinical and genetic data from 13,000 subjects. A key focus of the study is the role of milk fats in the diet, particularly low trans-fats and higher monounsaturated fat containing milks. There are also numerous dietary and clinical trials looking at the physiological effect of milk proteins, please see (Krissansen, 2007) and (Saito, 2008) for reviews of the current work. Microarray analysis of bovine, human and mouse mammary microarrays have been carried out (Suchyta et al., 2003a, Adjaye et al., 2004, Suchyta et al., 2003b, Lemkin et al., 2000, Stein et al., 2004, Stein et al., 2005, Kaminski et al., 2005, Clarkson and Watson, 2003). Lemay et al (Lemay et al., 2007) used bioinformatic approaches to examine gene regulation through the lactation process.

[1] http://lactoknow.ucdavis.edu/about-us
[2] http://www.ucd.ie/lipgene

Chapter 3 – Named Entity Recognition (NER) using MMTx

The first step of creating a domain related text-mining system requires a Named Entity Recognition (NER) system to identify biological entities within the scientific literature. Chapter 3 outlines an analysis of a particular system, MMTx which tags free text with biological concepts found within the UMLS Metathesaurus. The chapter then describes two key issues with MMTx and details methods to combat them.

A number of systems could have been used to perform this process such as Termino and NLProt, however, upon initial assessment their respective performance on milk-related literature was not adequate. A hand-written rule-based pattern matching script was also attempted but was quickly dropped in preference to MMTx due to its high degree of precision, biological scope and its ease of use and adaptability; features that were not afforded by other systems.

3.1 MetaMap transfer (MMTx) program

The performance of the MMTx program was evaluated to assess the validity of using it to perform NER in the *MilkMine* text-mining system. Although there are different running options available (for example, preference can be given to allow gaps in the name of the entity, ignore the order of words in the name or to prefer multiple concepts[1]) the default behaviour of MMTx is optimised for text-mining systems to give high recall and precision and therefore this was used for the evaluation.

3.2 MMTx data files

The MMTx program makes use of the 2005AC UMLS dataset (see Figure 3.9). Given the large size of the UMLS, only certain sources were selected (English only). The customised UMLS dataset improved performance to process 30 sentences per minute, a 12 fold improved performance compared with using the entire UMLS dataset. Although this performance is still slow, particularly when processing large sentence sets, this is primarily due to the heavy computational load which MMTx must deal with. Over 600,000 term variants were contained within the MMTx dataset which need to be matched against the text using the four metrics described in Section 2.3.4.

[1]See http://mmtx.nlm.nih.gov/mmtx.shtml for a full list of available running options.

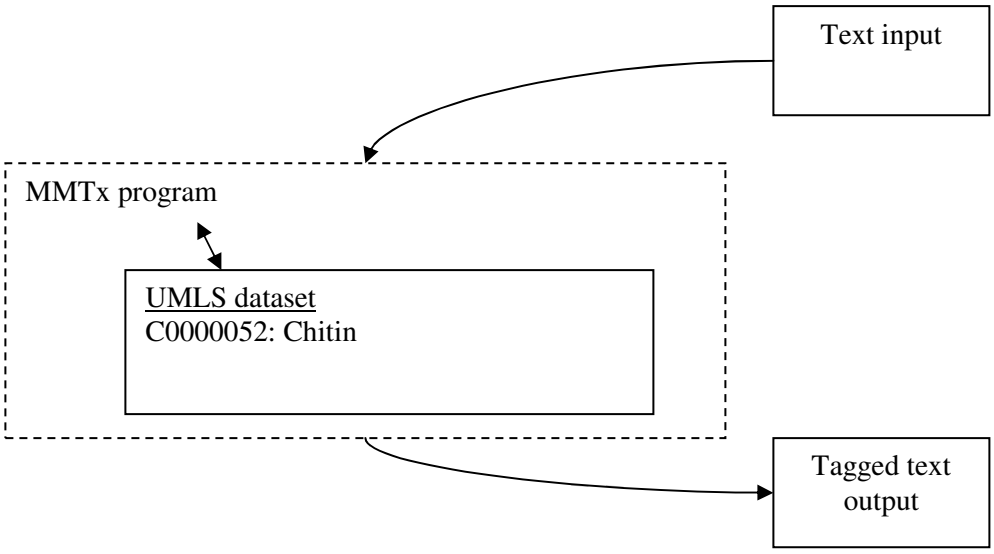


Figure 3.9: The MMTx program uses a UMLS dataset to tag text, comparing the text against the list of concepts in the UMLS it will assign the closest matches (see Section 2.4.3).

3.3 Evaluation of MMTx

To assess the recall and precision of the MMTx program against a known biological corpus the program (version 2.4.B) was used to tag an entire set of MeSH terms (MeSH 2005). The 2005AC UMLS dataset includes the MeSH 2005 and so maximum recall and precision were theoretically attainable. Any loss of recall of precision performance could then be directly attributable to the performance of the MMTx program.

3.3.1 MeSH 2005

The MeSH 2005 dataset was downloaded from the MeSH website and the preferred name[1] for each entry was extracted (22,861 distinct and preferred names). This list of terms was then analysed using the MMTx program. A term given a single UMLS candidate concept with an exact match score was assumed to be correct.

[1] A preferred name is a single name which has been chosen to be representative of all the names for that particular concept.

The MMTx tagging analysis results were as follows:

- 27,396 candidate concepts were identified (19.8% more than the number of MeSH terms in the list).
- 190 MeSH terms (0.8%) received no candidate concepts.
- 19,434 MeSH terms (85.0%) received a **single** candidate concept, of which 18,658 (96.0%) were exact matches (*i.e.* 81.6% of the total number of MeSH terms received a single, exact candidate concept match).
- 3,237 MeSH terms (14.2 %) received > 1 candidate concept, of which 2,765 (85.4%) had at least one exact candidate match (*i.e.* 12.1% of the total number of MeSH terms received an exact candidate concept match *and* at least one other lesser match).

In summary 21,432 MeSH terms (93.75%) received an exact match score and 18,658 (81.6%) received an exact match as the only concept, suggesting a recall of 0.93 and precision of 0.82.

3.3.2 MeSH 2006

The above analysis was repeated for the MeSH 2006 dataset containing 972 more terms to identify how the program reacted to new terminology.

The MMTx tagging analysis results were as follows:

- 29,012 candidate concepts were identified (21.48% more than the number of MeSH terms).
- 293 MeSH terms (1.2%) received no candidate concepts.
- 19,895 MeSH terms (83.3%) received a **single** candidate concept, of which 18,896 (95.0%) were exact matches (*i.e.* 79.1% of the total number of MeSH terms received a single, exact candidate concept match).
- 3,696 MeSH terms (15.5 %) received > 1 candidate concept, of which 2,824 (76.4%) had at least one exact candidate match (*i.e.* 11.8% of the total number of MeSH terms received an exact candidate concept match *and* at least one other lesser match).

In summary 21,721 MeSH terms (90.95%) received an exact match score and 18,896 (79.1%) received an exact match as the only concept, suggesting a recall of 0.91 and precision of 0.80.

In conclusion, the MeSH 2005 dataset was analysed more accurately than the 2006 as would be expected, however, 100% recall and precision were still not attained in either case. Errors causing the loss of performance were largely due to the phrase chunking part of the MMTx analysis of the text. Longer phrases are initially parsed into two separate phrases which are then treated separately by the program. Therefore an exact match cannot be found in these circumstances. Incorrect phrase chunking was also due to punctuation within the MeSH terms *e.g.* commas, brackets.

Another reason for the inaccurate or incomplete tagging was that only preferred names of the MeSH terms were used in the tagging process. While these should have been represented in the 2005AC UMLS dataset, use of secondary names of MeSH terms would probably have improved the recall of the terms, although this would have caused a decrease in precision.

3.4 False positive tagging analysis

MMTx was found to tag erroneously particular phrases, for example, the term ‘fat’ is tagged with the UMLS concept ‘FAT protein[Amino Acid, Peptide or Protein]’. A sampling process was used to identify automatically false positives which were consistently re-occurring.

A Perl script was written to calculate the document frequency for each concept in a large set of MMTx tagged citations (title and abstract sentences). A score was given to each concept (see Equation 1) and a cut-off value of 90,000 was selected where the number of citations tagged with a given concept is grossly overrepresented in comparison to a Medline search for that concept.

Score (S) = (C / M) * T

Equation 1

C = percentage of sentences that the concept is tagged to.

M = percentage of Medline citations that the concept is mentioned in (found by performing a Medline query using the preferred name of the concept).

T = number of citations that the concept is tagged to.

It is accepted that the document frequency produced by the MMTx tagging did not produce identical results to the frequency returned through Medline searches, however, the technique did successfully identify concepts with grossly disproportionate tagging frequency. For some key examples see Table 3.8.

Table 3.8: Key examples of false positive tagging frequency of free text by the MMTx program.

Concept Code	Concept	Reason for erroneous tagging
C1524028	Intraepithelial Neoplasia of the Mouse Mammary Gland	The concept has the synonym MIN and is incorrectly tagged to any sentence with the shortened form of the unit minute *i.e.* 'min'.
C1435181	FAT protein, human	Tagged to the term 'fat'.
C0221284	Leptocyte	A single Medline citation (Significance of the target cell (leptocyte) in peripheral blood smears…) led to 'target cell' becoming a synonym of the UMLS concept which is widely used in a general sense.

Sometimes the UMLS concept preferred name is not found in a natural language format (*e.g.* "Sampling - Surgical action") or are extended names (*e.g.* "Intraepithelial Neoplasia of the Mouse Mammary Gland") thus a Medline search may not be completely accurate although the search is not for a quoted phrase (exact match). Some UMLS concepts are tagged less than the Medline counts but have such a high average number of tags per sentence suggesting that there

is still a significant overtagging going on, therefore these were also included in the stop concept list. Since the script only removes those concepts which have an average citation occurrence way above the Medline occurrence, there is no risk of losing those concepts for which there is poor literature coverage in the storage files (except for the fact that false positives in this case may be missed due to the low number of sentences tagged).

Concepts identified as false positives are removed from the MMTx tagged data and subsequent analysis. Although there will be some correct tagging occurrences that are removed by this method, the benefit was deemed advantageous to the overall system.

3.5 Organism name extraction

Unfortunately MMTx cannot tag abbreviated organism names properly as the period character after the initial capital letter (*e.g. B. taurus*), causes the MMTx phrase chunker to put the name into two separate phrases. Therefore the name is not mapped, despite being present in the UMLS metathesaurus in abbreviated and full forms.

To rectify this error a pattern-matching perl script was written to extract all of the shortened organism names from a milk literature corpus (the results of a 'Milk[MH]' search in PubMed). A threshold of at least five occurrences in the 43875 citation corpus gave a list of 155 organisms. Another script replaces the shortened notations of these organisms with the full name and is used to process any literature which is stored in *MilkMine*. The full organism name notation is then correctly tagged by MMTx.

Chapter 4 – Towards a canonical representation of milk related terminology

This chapter outlines the process of generating a set of terminology which is associated with a particular domain of scientific literature, in this case milk and milk protein literature. Many text-mining systems use generic biological terminology sources to perform NER, for example EBIMed (Rebholz-Schuhmann, 2005) uses UniProt and Gene Ontology terms to identify proteins and processes in text. However, although a generic system has the advantage of being generally applicable to a wide variety of subject areas (or domains), it may fail to recognise the terminological subtleties within a specific subject area. Furthermore, the system may make incorrect identification of concepts that are of high importance within the domain.

To improve the performance of text-mining systems for a particular niche of biology such as protein-protein interactions, developers often manually create lists of keywords or stop words. For example Swanson collated a list of stop words for his ARROWSMITH system (Smalheiser and Swanson, 1998). However, this process requires a huge input of time and experience to discern and record all of the terminology that is important and relevant to a particular domain. For the text-mining system to be transferred to another domain the effort must be repeated, thus massively reducing the potential applicability of the system. One of the key aims of the *MilkMine* project was to create a method to automatically identify domain (*e.g.* milk and milk protein) related terminology from scientific literature in order to improve the recall and precision of a subsequent text-mining algorithm.

4.1 Selection of a corpus of milk related literature

A corpus of milk related literature was retrieved from the Medline database using searches with several milk related Medical Subject Headings (see Table 4.9). Four of these MeSH terms (Milk; Milk protein; Lactation and Colostrum) returned the vast majority of the publications identified (99.6%) and these were selected as key MeSH terms[1] for the retrieval of milk related literature (see Figure 4.10). Although there is some overlap between these four sub-literatures within the milk related literature corpus, 86% of the publications are distinct to only one of the sub-literatures. Therefore selection of these four MeSH terms provides a high recall and broad coverage of milk related literature. Citations were downloaded in Medline format, a format

[1] From this point on these four MeSH terms will be referred to as the 'key milk related MeSH terms'.

which includes the title, abstract, MeSH and substance terms of the article (see Appendix D for an example of this format).

Table 4.9: Searching Medline for milk related literature using selected MeSH headings[1].

Search Number	Search Term	Publications		
		Total	In English	With abstracts
1	**Milk [MH]**	**43270**	**35474**	**24695**
2	**Lactation [MH]**	**41686**	**35863**	**25551**
3	**Milk protein [MH]**	**22802**	**20398**	**16203**
4	Breast feeding [MH] *(subset of search 2)*	17465	14549	9708
5	Milk, Human [MH] *(subset of search 1)*	11748	9513	6806
6	**Colostrum [MH]**	**4302**	**3675**	**2564**
7	Milk Hypersensitivity [MH]	968	804	712
8	Whey protein [Substance]	824	804	786
9	Cultured milk products [MH] *(subset of search 1)*	805	714	679
10	Milk substitutes [MH] *(subset of search 1)*	510	481	387
11	Infant formula [MH] *(subset of search 1)*	443	419	325
12	Milk Ejection [MH] *(subset of search 2)*	442	397	337
13	Whey acidic proteins [Substance]	213	211	202
14	Milk fat globule [Substance]	78	78	71
15	Soy milk [MH]	73	68	64
16	Mammary-derived growth factor 1 [Substance]	6	6	5
17	Milk-derived factor [Substance]	1	1	1
	TOTAL	**95,601**	**80,612**	**57,561**

[1] Searches were made on 30th May 2006 through the PubMed interface (www.pubmed.gov).

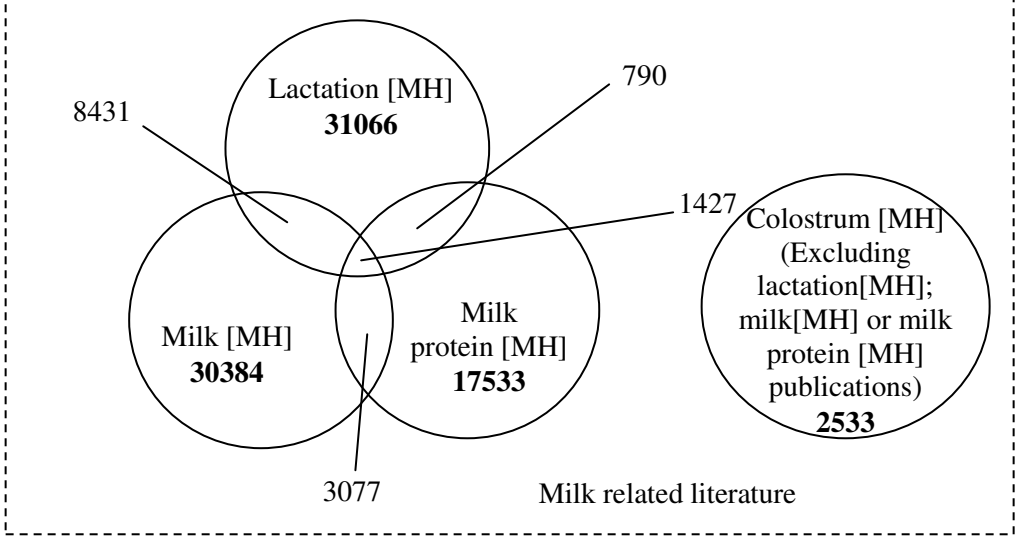


Figure 4.10: Retrieval of milk related literature from Medline by use of four key milk related MeSH terms (Numbers shown are number of citations in Medline, searches were made on 30th May 2006 through the PubMed interface, www.pubmed.gov).

MeSH terms are assigned to publications in Medline by human annotators therefore we can have confidence that the corpus retrieved contains significant information relating to milk. However as MeSH terms are assigned based on the full text of a publication it does not necessarily follow that the term will be mentioned in the actual abstract. Also due to the hierarchical structure of the MeSH vocabulary the PubMed search will be expanded to include child terms of that MeSH term (unless the user specifically turns this function off). For example, searching Medline with the MeSH term Milk[MH], will be expanded to search for literature of the child terms: Cultured milk products or Infant formula (see Figure 4.11).

However, while these issues may mean that citations retrieved are not necessarily about milk (*e.g.* they may include publications about yoghurt) this is not detrimental to retrieving a corpus of milk related literature since it will include publications on the periphery of the domain, publications which may contain key connections to other biological domains.

Fluids and Secretions [A12]
- Bodily Secretions [A12.200]
  - Milk [A12.200.455]
  - Milk, Human [A12.200.467]

Food and Beverages [J02]
- Beverages [J02.200]
  - Milk [J02.200.700]
    - Cultured Milk Products [J02.200.700.124]
    - Infant Formula [J02.200.700.249]
    - Milk, Human [J02.200.700.500]
  - Milk Substitutes [J02.200.712]

- Food [J02.500]
  - Dairy Products [J02.500.350]
    - Milk [J02.500.350.525]
      - Cultured Milk Products [J02.500.350.525.221]
      - Infant Formula [J02.500.350.525.332]
      - Milk, Human [J02.500.350.525.500]
      - Milk Protein [J02.500.350.525.520]

Figure 4.11: Hierarchical structure of MeSH showing the parent and child terms of the MeSH term Milk (MeSH hierarchy codes are displayed in square brackets).

4.2 Identification of milk related terminology

Having identified and retrieved a milk related corpus of peer-reviewed publications, terminology related to that corpus could then be identified. This was achieved through manual and automatic analysis methods (see Figure 4.12).

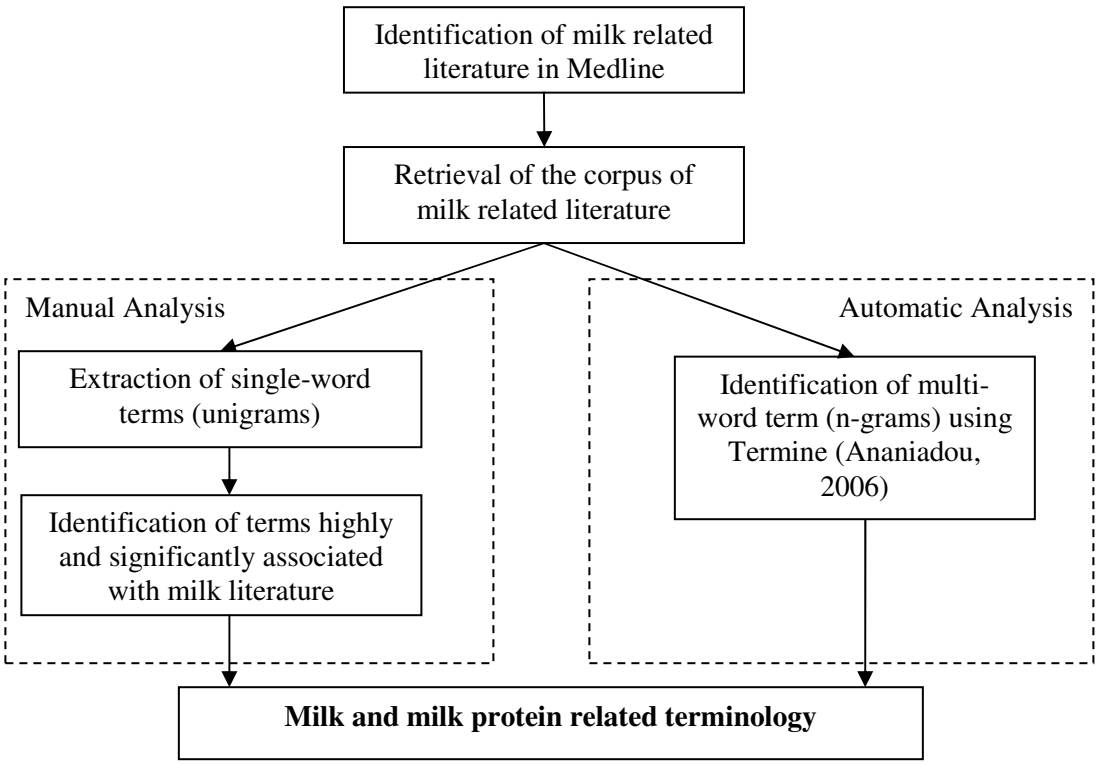


Figure 4.12: The identification of single-word (unigram) and multi-word (n-gram) terminology related to a particular domain of scientific literature, for example literature describing milk and milk proteins.

4.2.1 Literature Gradient Technique (LGT)

To identify terminology that was particularly associated with milk literature, a series of corpora of increasingly milk related literature were created, from general scientific literature to very milk related literature. By identifying terminology that was found in milk related literature much more frequently than in general literature, milk related terminology was determined. This technique has been named the literature gradient technique (LGT).

4.2.1.1 Creation of a series of corpora for LGT

A set of rules was devised to define 6 categories of milk related literature as shown in Table 4.10 and these were use to create Medline searches. The first 8,500 publications with abstracts available (by most recent) were retrieved for each corpus. Titles and abstract sentences were extracted and stored locally for terminological analysis.

The 'Very milk related' category was defined as containing publications assigned with a key milk related MeSH term (*i.e.* Milk; Milk protein; Lactation or Colostrum) where the term represented the main focus of the publication[1]. A 'Highly milk related' category contained publications assigned with a key milk related MeSH term with no stipulation that this must be the main thrust of the article.

Publications for the 'Milk related', 'Slightly milk related' and 'Vaguely milk related' corpora were collected from journals which publish a certain proportion of milk related articles respectively (see Table 4.10).

To identify a list of milk related journals, a Milk[MH] search was performed in Medline and a list of the publishing was extracted. Each journal in this list was given a score using Equation 2. For example, the Journal of Dairy Research had published 1298 articles in Medline, 751 of which were annotated with the MeSH term Milk[MH]; thus the journal had a MRV of 58%. The journal list was then ranked by MRV score. To negate a preference towards journals with a low number of publications, any journal with less than 1,000 publications in Medline or less than 100 Milk[MH] publications were removed, leaving a list of 2000 journals and 72 journals respectively (see Appendix E).

$$\textbf{Milk Related Value of Journal (MRV)} = \frac{\text{Number of publications with Milk[MH] MeSH term from Journal A}}{\text{Total number of publications published by Journal A}} \times 100$$

Equation 2

The final category was simply a random set of biological publications, which could be used to represent the terminological baseline against which terminology from the other categories could be compared.

[1] In the Medline database MeSH terms which represent the main focus of the publication are marked with an asterisk (see Appendix D). A search restriction can be placed on a PubMed search to retrieve only citations which have a particular main MeSH term (*e.g.* milk[MAJR]).

The MRV analysis was repeated for the other key milk related MeSH terms (Milk Protein; Lactation and Colostrum); however the colostrum related corpus only produced 5 journals passing the criteria specified (1,000 publications in Medline and 100 Colostrum[MH] publications), therefore subsequent analysis was not performed for this MeSH term. Thus the total number of corpora retrieved for terminological analysis was 18 (6 for Milk[MH] literature, 6 for Milk protein[MH] literature and 6 for Lactation[MH] literature, see Figure 4.13).

Table 4.10: Categories of milk related literature in Medline.

Milk related category	Category code	Description	PubMed Search*	Number of Publications (with abstracts)
Very milk related	5	Publications where milk is the main focus of the article.	Milk[MAJR]	23,122 (14,513)
Highly milk related	4	Publications where milk is a key theme of the article.	Milk[MH]	39,586 (24,517)
Milk related	3	Publications from the most milk related journals (*i.e.* journals which had an MRV above 10%).	(List of appropriate journals†)	10,608 (8,667)
Slightly milk related	2	Publications from semi milk related journals (*i.e.* journals which had an MRV between 10% and 5%).	(List of appropriate journals†)	59,624 (41,839)
Vaguely milk related	1	Publications from low milk related journals (*i.e.* journals which had an MRV of less than 5%).	(List of appropriate journals†)	948,729 (501,587)
General Literature	0	Random selection of publications	Medline[sb]	13,716,333 (7,848,919)

*Searches were made on 2nd May 2006. †See Appendix E for the list of journals used.

The milk related value of each retrieved category was verified by calculating the MRV based on the downloaded corpora and comparing them to the theoretical MRV. The values obtained were found to be within the ranges required for LGT (see Table 4.10). The MeSH term frequency in the downloaded citations is shown in Table 4.11.

Table 4.11: Verification of the milk related value of a series of corpora retrieved from Medline.

Milk related category	Number of citations with Milk[MH]	% of the corpus with Milk[MH]	# citations with Milk[MH] or child term	% of the corpus with Milk[MH] or child term	Target value
5	4778.33	56.22	8500.00	100.00	100%
4	4951.67	58.25	8500.00	100.00	100%
3	2778.33	32.69	2970.67	34.95	>10%
2	284.67	3.35	484.67	5.70	between 10 and 5%
1	53.67	0.63	99.00	1.16	< 5%
0	11.67	0.14	23.33	0.27	0%

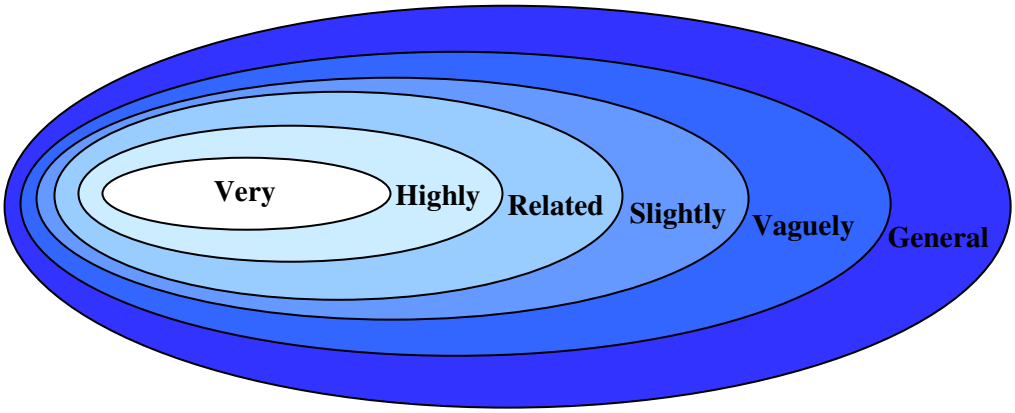


Figure 4.13: Creation of a series of literature corpora with an increasing level of association with milk for each of the key milk related MeSH terms (Milk; Milk protein; Lactation).

4.2.1.2 Manual identification of milk related terminology (unigrams)

Having defined and retrieved these 18 corpora, manual and automated terminological extraction could be performed. Unigram terms (*i.e.* single words) were extracted from the titles and abstracts of each corpora using a Perl script written by the author. Unigrams were defined as being those words which are broken by a word boundary[1] with the exception of hyphens, decimal points and apostrophes. For example "alpha-casein" was treated as a unigram term while "alpha casein" was not. These terms were converted to lower case and reduced to their word stem using the Porter stemmer (*e.g.* 'activation' becomes 'activat') . This additional

[1] A word boundary is any non-alphanumeric character, for example a space, bracket or semi-colon.

processing reduces redundancy in the term lists without a significant loss of data, for example 'Milks' and 'milk' are collapsed to same term 'milk'. Any terms that contained only numeric and/or period (.), hyphen (-) or apostrophe (') characters (for example "1-25") were removed from the list of terms.

The resulting term list was normalised to express the frequency of each term as a percentage of the total number of terms within the corpus. This negated the small amount of variation between the total numbers of terms in each corpus. The terms were then ranked by normalised frequency within the corpora (see Table 4.12).

Table 4.12: **Single word term extraction (lower case, stemmed) from milk related literature. Figures shown are average values calculated for the key milk related MeSH terms (Milk; Milk protein; Lactation) and using 8,500 publications in each corpus (individual results are shown in Appendix F).**

Milk related category	Number of journals represented	Total number of unigrams	Number of unique unigrams	Percentage unique over total unigrams
Very related (5)	1230	1741220	33149	1.90
Highly related (4)	1245	1872900	36906	1.97
Related (3)	2	1807370	24657	1.36
Slightly related (2)	7	1923142	35206	1.83
Vaguely related (1)	30	1799290	56753	3.15
General literature (0)	639	1682919	50459	3.00
Average	**451**	**1546692**	**33876**	**2.74**

As would be expected, there was a relatively small variation[1] (13.3%) between the total numbers of unigrams in each corpus, however there is a huge variation (81.2%) in the number of unique terms. This large variation can be explained by two factors. Firstly the narrowness of the subject area of the corpus; searching for a particular topic will produce a restricted terminology while a more general search will cover a far broader range of topics (Swanson et al., 2006). Therefore as the 'Very milk related' corpus (5) is the most restrictive search space (see Figure 4.13), it has a higher redundancy in comparison to general literature. Category 4 is less restrictive and thus produces less redundant terminology. Although categories 1 and 0 have fewer journals

[1] Percentage variation = (maximum difference / total number of terms) * 100

represented in comparison to categories 5 and 4, the unrestricted search space has resulted in a decrease in term redundancy (see Figure 4.14).

The second factor affecting terminological redundancy within a corpus is the variation in the number of journals used. The style, format and flavour of language of publications within a given journal are often similar. The 'Related' corpus (3) is drawn from just two journals, therefore the highest level of redundancy was found in this corpus. In comparison to categories 5 and 4 (1230 and 1245 journals respectively, see Table 4.12), category 3 is much more redundant. This analysis shows that the variation in terminology within a domain (*i.e.* milk, see categories 5 & 4) is greater than within a specific journal (category 3, see Figure 4.14).

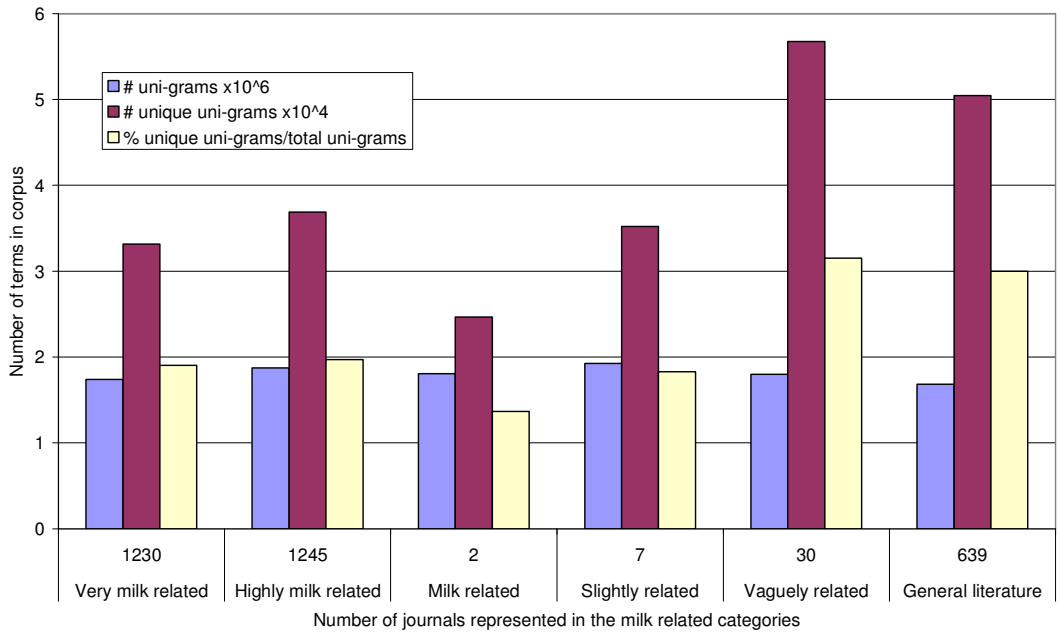


Figure 4.14: Redundancy of unigrams against number of journals within a gradient of milk related literature corpora.

Having shown that non-redundant milk related terminology was present in categories 5 and 4, these corpora were used in the following steps. Initially terms that were found in equal measure in both milk (categories 5 and 4) and general literature (*i.e.* stop words) were removed. Milk related terminology was then identified by comparing the unigram term list from the milk related category with that of the general corpora (0). Terms which have a normalised frequency score in

the milk related literature of at least 10-fold higher than in general literature. This analysis produced 628 unigrams including casein, whey and oligosaccharides.

4.2.2 Automated analysis of milk related literature using Termine

To identify milk related multi-word terms (n-grams), the Termine (Okazaki and Ananiadou, 2006, Ananiadou, 2006) application was used on the four key milk related MeSH term literatures: Milk; Milk protein; Lactation and Colostrum. The Termine programme assesses the significance of a term (C-value) within a corpus based on four factors:

1. the frequency of a candidate term in the corpus
2. the frequency of the candidate term as a part of other longer candidate terms
3. the total number of longer candidate terms and
4. the length of the candidate term.

4.2.2.1 Effect of corpus size on term significance

The effect of corpus size on the precision and recall of milk related terminology using Termine was analysed. The top 3 terms from the Milk[MH] Termine analysis of 1,250 citations were used to perform a series of analyses for 50 – 1250 citations. An analysis was performed to identify the minimum number of citations required to ensure that n-grams produced from the corpus were significantly domain related[1].

The absolute C-value was found to rise proportionately with the corpus size *i.e.* the C-value does not reach a maximum and will increase corresponding to the citation set size (see Figure 4.15). Therefore using a huge citation set may not produce higher precision for the top milk related terms, however recall will be better. Even a low corpus size (above 250 citations) gave a good representation of the top terms.

[1] Although the maximum capacity of the Termine web program was 1,250 citations, automated analysis was also completed on a 'Milk[MH]' corpus containing 8,500 publications by kind request to the Termine developers.

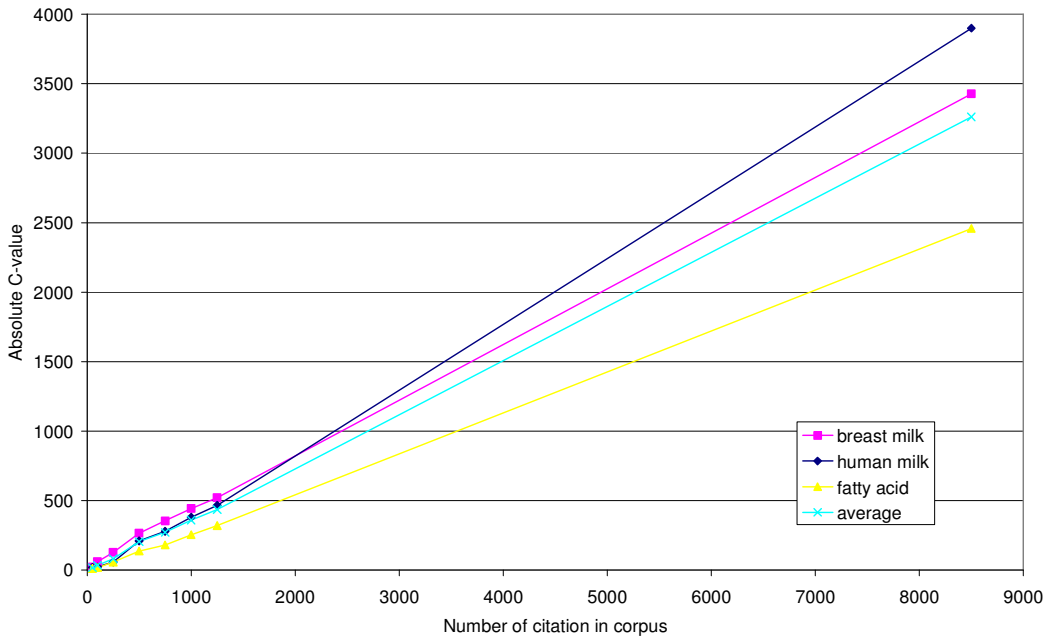


Figure 4.15: Linear relationship between absolute C-value and corpus size using the Termine term extraction program.

To assess whether the number of citations made a difference to the relative C-value [term C-value/total C-value] given to a term, the milk[MH] corpus was analysed (*i.e.* does using a bigger citation set give you a more definite term).The significance of the C-value of a term within a corpus levels out at approximately 750 citations *i.e.* the significance of the top terms does not change as the citation size increases further (see Figure 4.16). Therefore, using at least 750 citations for each Medline search allows us to extract domain related terminology with a reasonable amount of significance. In the analysis, either 1,250 was used and was therefore well into the significant zone.

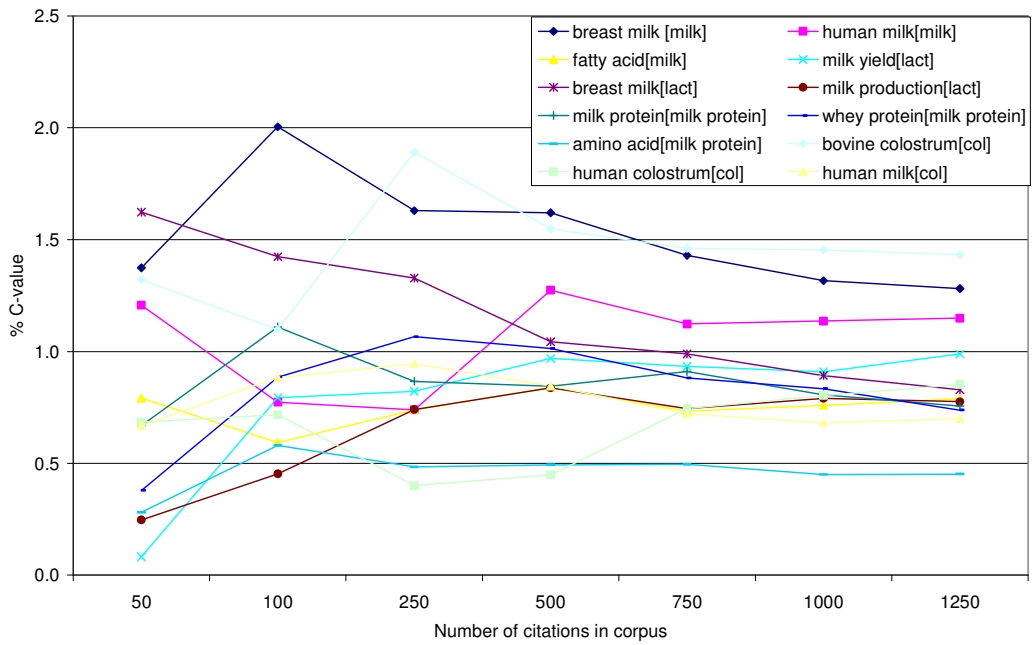


Figure 4.16: Relative C-value (term C-value/total C-value) against corpus size for key milk related MeSH term literature (Milk, Lactation, Milk protein and Colostrum[MH]), using the top three terms from each 1250 corpus.

4.2.3 Results of domain related terminological analysis

The top 500 unigrams and n-grams (see Sections 4.2.1.2 and 4.2.2 respectively) from each list (Milk, Milk proteins, Lactation, Colostrum) and were taken and combined to create a non-redundant list of milk related terminology. Unigrams which had been stemmed were re-expanded manually and where these were part of multi-word terms (n-grams), they were replaced with the full term. For example, baylei was replaced with Bayley Scales.

In total, 3,611 terms were identified from the terminological analysis of milk related literature. 512 of these were not currently available in the UMLS metathesaurus; 750 were new variants or spellings of existing concepts and 2,394 were already present in the 2006AB_custom UMLS dataset.

The technique was found to perform well at identifying milk related terminology, including milk related acronyms, for example HM (human milk) or LBW (low birth weight). Milk related experimental procedures were also identified, such as lactose-maldigestion test or the CHARM

test as well as milk related institutions such as the FAO (Food and Agriculture Organisation) and the NCHS (National Centre for Health Statistics). Dairy techniques and processes, for example Pretoria pasteurization were also found, as were important species such as the common milk product bacterial strain *Lactobacillus acidophilus* and the cow breeds such as Holstein.

4.2.4 Customisation of the UMLS using MetamorphoSys

As with the 2005AC UMLS dataset used for the Named Entity Recognition (NER) analysis (see Chapter 3), the 2006AB UMLS dataset was customised using MetamorphoSys to select a list of suitable sources list (see Appendix G). At this point we have a non-specific ontological resource (see Figure 4.17a) based on the source vocabularies we have included in our UMLS dataset.

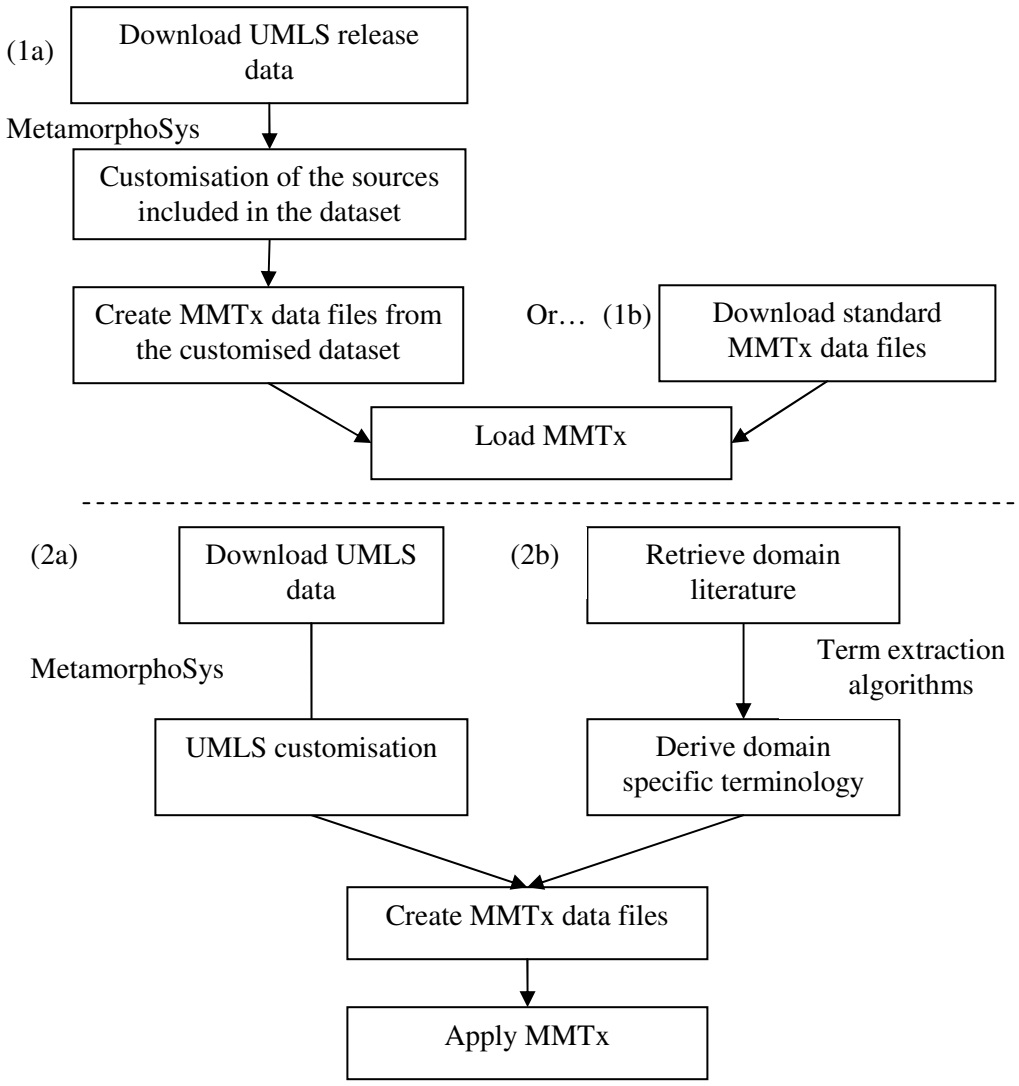


Figure 4.17: 1 - Standard procedure either through customisation of the entire dataset (1a) or the use of a preprepared filtered dataset (1b); **2 -** *MilkMine* procedure: Extra steps were added to the standard procedure to improve the coverage of milk related terminology in the UMLS dataset such to improve precision and recall of the subsequent text-mining algorithm.

4.2.5 Complementation of the UMLS dataset

There are four data options from which to operate MMTx with (see Figure 4.17):

1. a standard UMLS dataset which can be downloaded from the UMLS server
2. create a customized UMLS dataset
3. create a custom dataset or

4. create a combined dataset from 2 and 3.

MilkMine used the fourth option to complement the customised 2006AB UMLS dataset with the milk related terminology identified in previous sections.

The identified milk related terminology (see Section 4.2.3) was tagged using MMTx to check whether there was adequate representation of each identified milk term in the 2006AB customised UMLS dataset. The MMTx tagging results were checked categorised into: correct recognition; partial recognition or incorrect recognition. Partial recognition included terms which had been correctly tagged in name but not in concept, for example, "calf" was tagged as Veal [Food] or Calf (Structure of calf of leg) [Body Location or Region]. Other examples of partial recognition were split into fragments, for example "California mastitis test" was tagged as California [Geographic Area]; mastitis [Disease or Syndrome] and Test (Testing) [Research Activity]. The term "Casein micelle" was tagged as Casein (Caseins) [Amino Acid, Peptide, or Protein, Biologically Active Substance] and Micelle (Micelles) [Substance].

Some of the incorrectly tagged terms were due to the fact that during the Termine process, all of the textual characters were reduced to lower case and as MMTx takes into account the case of a term when it calculates the term likelihood, terms such as IgG, lost much of their inherent topology and therefore were not recognised by MMTx.

The system was also successful in finding more obscure but important milk related terminology such as mother-infant dyad (mother and infant thought of as one unit). Infrequent milk related abbreviations were not used to complement the UMLS since this may actually reduce the precision of the system as it may be mapped inaccurately to other common concepts thus adding noise to the MMTx output.

Some terms were slightly surprising by their presence in the milk related term list such as 'main outcome' which you would expect to be domain independent. The term appeared since the format 'MAIN OUTCOME MEASURES:' was used in a particular milk related journal but was not abundant in the general literature. This example shows that the specific style of certain journal (even within abstracts) can influence the term extraction of the system.

The 512 identified terms which were not found in the 2006AB_custom UMLS dataset were assigned custom identifiers and semantic types. These terms were then included in the metathesaurus and the MMTx datafiles were created using the scripts supplied with the program. This step filters the raw data, creates all of the necessary indexes and files for running MMTx.

At this point we now have a general biological ontological resource which is enriched with terminology related to a particular biological domain (*i.e.* milk).

Chapter 5 - Co-occurrence Relation Extraction Algorithm

To mine the vast amount of scientific literature, a text-mining algorithm was created. This was based on the principle of co-occurrence relation extraction whereby if two entities are co-located in the same piece of text then this indicates a potential relationship between them (see Figure 1.2). The algorithm has two main constituent parts (see Figure 5.18):

There are three main parts to the text-mining program:

1. Retrieval and storage script. This script searches Medline and retrieves the results before extracting and tagging all the sentences with the 2006AB_custom UMLS dataset (see Chapter 4) using the MMTx program. The sentences and concept mappings are then stored locally. This script allows for expansion of the database literature coverage without doing any A-B-C calculations, thus is useful for database updates and simple expansion to other domains.
2. Literature relation extraction analysis script. This script takes a user query (*i.e.* A term for open discovery method or A and C terms for closed discovery), produces a set of intermediate (B) concepts and filters the B-concept list. From the intermediate concept list, the algorithm generates and filters the C concept list. Finally the script filters and ranks the A-C list.

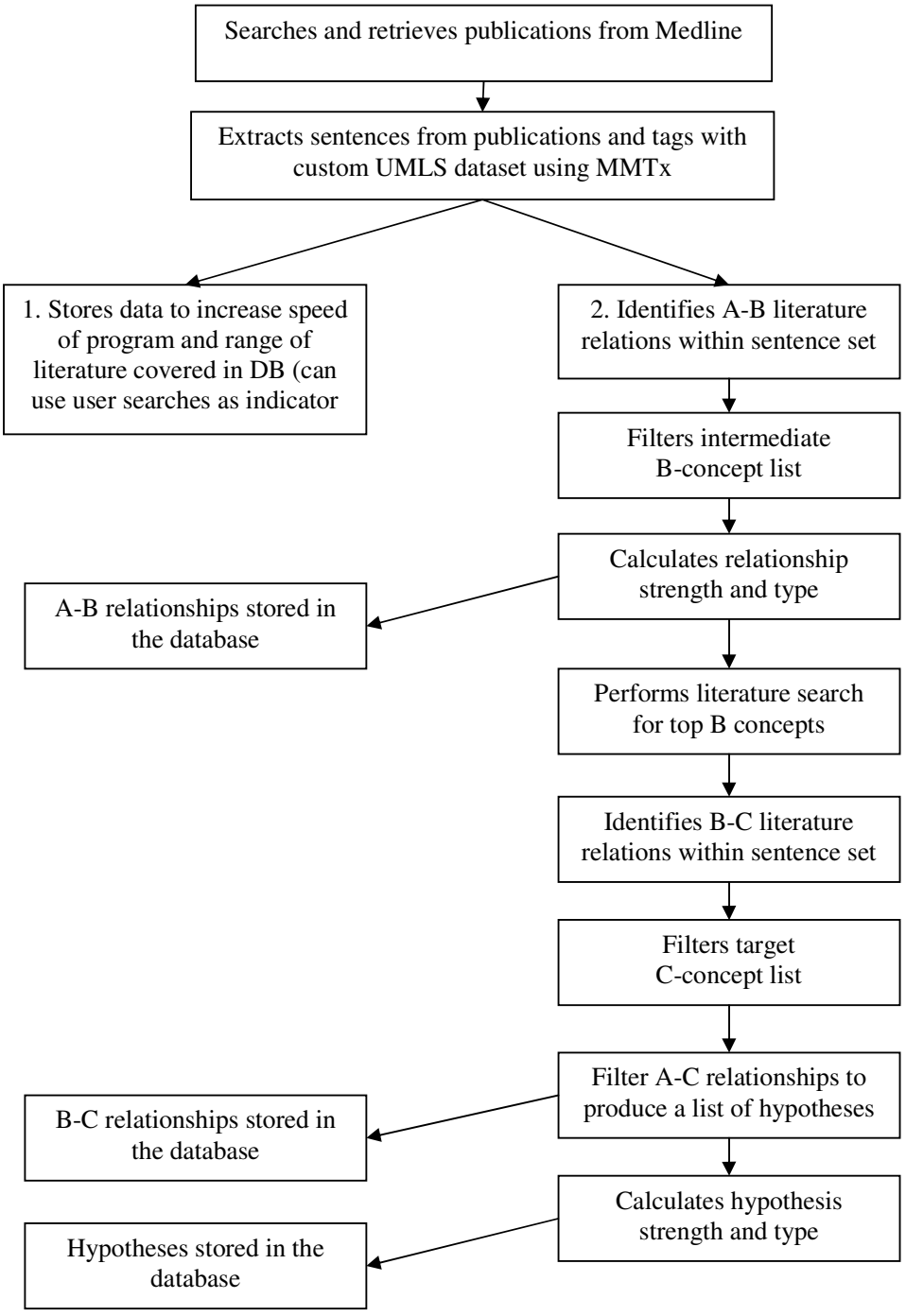


Figure 5.18: Overview of the *MilkMine* text-mining algorithm structure.

5.1 Document retrieval

A query term is given by the user. This is then expanded using the UMLS to include the synonyms of the query term. The expanded query term is then sent to Medline using the PubMed e-utilities facility. A list of PubMed IDs are retrieved and checked against the locally stored literature database. Any citations that have not been previously stored are downloaded in Medline file format (see Appendix D). A limit of 1,000 publications was placed on the document retrieval so that the relationship analysis did not become excessively large. It was assumed that the relationship significance will balance out in a similar manner to the term significance found (*i.e.* above 750 citations, see Section 4.2.2.1).

5.2 Sentence extraction and Named Entity Recognition (NER)

The downloaded publications were then parsed to extract the free text (*i.e.* title and abstract), MeSH and substance terms and the citation information. This information was stored locally. The MedPost tagger (Smith et al., 2004) was used to identify sentence boundaries within the free text sections. Each sentence was extracted and given unique ID. The sentence set was passed through the MMTx program to tag UMLS concepts from the customised 2006AB UMLS dataset (see Chapter 4).

Sentences containing UMLS concepts belonging to semantic types identified as being applicable to text-mining systems are extracted for literature relation analysis.

5.3 Intermediate (B) concept list reduction

The intermediate (B) concept list was reduced by applying two categories of filter: term based and relation-based filters.

5.3.1 Term-based filters

Term-based filters remove intermediate terms purely on the attributes of the B-concept itself.

5.3.1.1 Stop concepts

Stop lists of terms that are naturally abundant within general language (*e.g.* ‘and’, ‘or’) are typically used in text-mining systems. One of the main problems with stop lists is their applicability to any given application. For example, the Arrowsmith project created their own

stop list based on manual curation. While manual curation gives a good accuracy, it requires a large amount of time and effort, therefore ways of automatically generating a list of stop terms or concepts is preferred. The MilkMine stop list was generated semi-automatically using the following four steps (see Table 5.13):

1. A stop list from NCBI was string matched against the MRCON file to generate a list of stop concepts[1] (*e.g.* And (C1515981); Or (C1518602); Not (C1518422)).
2. A number of irrelevant semantic types (*e.g.* Professional society) were labelled as stop types.
3. Overabundant concepts within a large sample of 8.7 million tagged sentences were identified. Any concept which appears more than 250,000 times in the literature set of 8,766,121 sentences *i.e.* 2.85 % of the sentence set (or average of at least once every 35 sentences). This analysis resulted in only a small set of stop concepts due to the fact that the UMLS is a biomedical thesaurus and therefore contains very few natural language stop concepts.
4. Additionally, further stop concepts are added manually.

In total the stop concept analysis produced a stop list of 7,886 distinct UMLS concepts (see Table 5.13). These stop concepts were removed from the tagged sentence data. 37,654,556 tagged concept instances were removed from 8,766,121 sentences, a 28.0% reduction in the number of concept tags. Applying both the stop concepts filter *and* the stop semantic type filter results in tagged concept instances being removed from the tagged sentence set, out of a total of 134,385,017 concept tags from 8,766,121 sentences (28.02% reduction).

Table 5.13: Stop concepts identified by various methods for the *MilkMine* application.

Source	Number of concepts	Found in UMLS concepts
NCBI Medline stop words	132	22
Stop concepts by semantic type	7816	7816
Overabundant stop concepts	37	37
Manually added stop concepts	24	24
TOTAL	**8009**	**7886**

[1] http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=stopwords

5.3.1.2 Hypothesis generation semantic types

A number of UMLS semantic types have been identified as being relevant for hypothesis generation in other biological systems (see Appendix H). These discovery types are classed as primary concepts. In the literature-mining analysis only Primary-Primary or Primary-Secondary concept relations are considered. This means that any sentences with no Primary concepts or with only one concept can be removed from the analysis.

5.3.1.3 General term filter

A straightforward way to reduce the B-term list is to remove general concepts. This is done using the MRHIER.RRF file of the customised UMLS dataset 2006AB_custom (see Chapter 4) which contains hierarchical information from the root of the source vocabulary to the concept (if this is relevant and available). 678,146 / 1,050,688 (64.5%) concepts from the 2006AB custom UMLS dataset include hierarchy data from their source vocabulary.

However, there is an inherent problem with using the hierarchy data as the variability of the source hierarchies is maintained within the UMLS. Although this is beneficial in some regards, it makes it difficult to apply a single hierarchical filter across the board, without potentially reducing the recall of the system. For example MeSH has four levels of generality while the Common Terminology Criteria for Adverse Events source only has one. An analysis was performed to identify the depth of generality within each source vocabulary. General terms are then removed from the subsequent literature relation analysis according to the source vocabulary the concepts came from.

5.3.1.4 Maximum and minimum document frequency in Medline

A document frequency filter removes any concept which has a document frequency in Medline above or below certain thresholds. Each concept in the 2006AB_custom UMLS dataset belonging to the discovery semantic type groups was searched against PubMed using the e-utilities facility. The preferred name for each concept only was used to reduce the search time overhead and the load placed on the PubMed server.

The maximum frequency was set to 15,000 and the minimum was set to 1 (*i.e.* the concept must return more than zero articles). This removes any concepts which are either not present in the

Medline database or are highly abundant and will therefore produce an excessive number of literature relations.

5.3.1.5 False positive concepts

Some concepts are erroneously tagged to sentences as false positives (as described in Section 3.4). The list of false positive tags generated by the NER analysis were added to the stop concept list.

5.3.2 Relation-based filters

Relation-based filters remove intermediate (B) concepts based upon relationships of the concepts within their respective source vocabularies or within the citation set in which they are located.

5.3.2.1 Parent/child filter

The parent/child filter removes any terms that are directly related to the user query term, based on the hierarchical data in the MRHIER file. If a concept has multiple hierarchies and one of them is a parent or child of the search term then the concept is removed from the B-list. For example the hierarchical relationship, Activin A *is_a* Activins is not an interesting literature relationship and so is removed from the B-term list.

5.3.2.2 Level of support

The level of support filter examines the support for a given literature relationship. Literature relations are ranked by number of sentences containing the relationship. Maximum and minimum values were set as 250 and 2 respectively. This filter removes literature relations above 250 which are deemed as being too well known to produce novel hypotheses, while relationships identified from only one or two sentences were removed to reduce the level of false positive relationships.

5.3.2.3 Concept hub connections

Biological networks are typically composed of hub and spoke concepts where a few concepts (hubs) will be linked to many other concepts (spokes) (He and Zhang, 2006). Although hubs are useful concepts, in text-mining analysis, it is often the lesser known relationships that are of more value. The huge number of relationships from the hub can drown out these more interesting relationships and therefore reduce the efficacy of the system. Consequently, a maximum allowable number of connections for each concept was set to 500. A minimum level was also set at 3 to reduce the number of false positive relationships and thus improve precision.

5.3.3 Hypothesis filters

Additional filters are placed on the hypotheses generated by the *MilkMine* algorithm. A maximum and minimum number of literature relations for a given hypothesis are set as 10 and 3 respectively. This ensures some novelty of the hypothesis while ensuring reasonable evidence to back it up.

5.3.4 Literature relationship categorisation

Literature relationship are categorised into five interaction types (Enzymatic interaction; Permanent interaction; Physical interaction; Positional interaction; Regulatory interaction) according to the presence of key interaction verbs within the sentences on which the relationship is based.

An interaction verb list was compiled from those created for (Hoffmann and Valencia, 2005, Chen and Sharp, 2004) and a list of UMLS concepts labelled with the semantic type 'Functional Concept'. This resulted in a list of 130 interaction verbs, including milk related verbs such as suckle, secrete and involute. The interaction type categories were assigned for each verb and this list was used to assign the interaction types for the literature relations. The most common interaction type is selected as the main interaction type for that relationship.

The strength of the interaction is also calculated as shown in Equation 3:

Relationship strength = number of sentences with link * (number of citations with link)^2

Equation 3

Hypotheses are categorised into the five interaction types (Enzymatic interaction; Permanent interaction; Physical interaction; Positional interaction; Regulatory interaction) according to the most abundant interaction type of the literature relations which have been used to generate that hypothesis. For example, if a hypothesis includes 8 Positional literature relations and 3 Physical literature relations, the hypothesis will be categorised as a positional hypothesis.

Chapter 6 – Creation and curation of the *MilkMine* database

Early on in the *MilkMine* project it was realized that there was a great potential to integrate biological data on milk proteins with not just milk-related literature but also the results of literature-mining analysis (see Chapter 4). This chapter outlines the process of creating the database and a description of its features (see Figure 6.19).

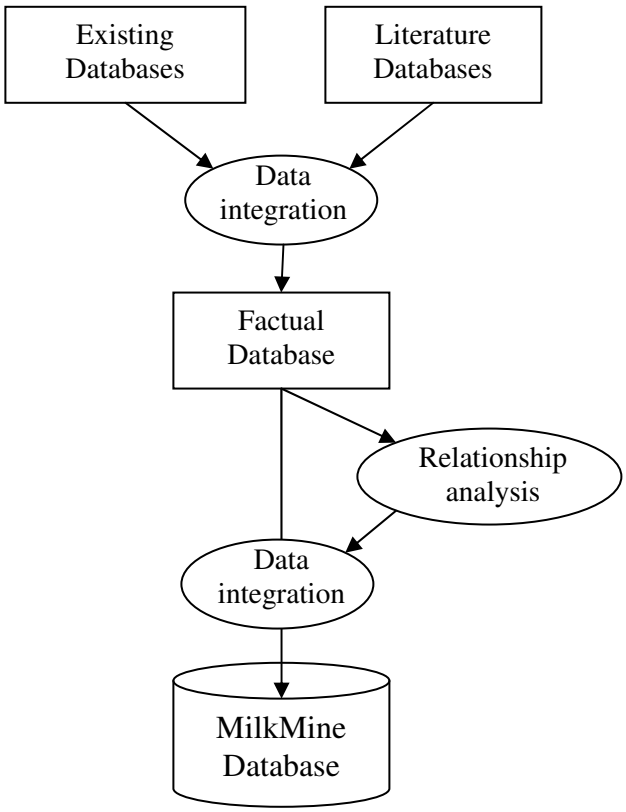


Figure 6.19: Information flow into the *MilkMine* database. Biological and literature data are downloaded from external sources and are integrated to create an initial factual database. Literature data is then statistically analysed to identify significant biological relationships which are in turn integrated with the factual database to create the *MilkMine* database.

6.1 Initial milkER (milk Extraction Resource) system

The first version of the database, the *milkER* database (Edwards et al., 2005) was designed and written by the author. This was based on BioSQL and BioPerl (an open source project to create a set of programs and scripts specifically for creating databases of biological data) and was

searchable using an application called DBSight. Data from a variety of sources were loaded into the *milkER* database including milk protein and bioactive peptide data, milk gene data and a small set of milk related sentences (50,000 sentences). These sentences were tagged with biological concepts using MMTx as described in Section 5.4.

The *milkER* web interface (see Figure 6.20) included functions such as simple and advanced search forms for proteins, genes or bioactive peptides (by name, identifier or sequence) as well as pages for FAQ, related links and a tutorial on effective searching in *milkER*. There were a number of drawbacks to this version:

1. Data integration in the database and the query interface were manually hard coded and were only moderately effective.
2. The size of the database had to be scaled up dramatically and would exceed the capacity of the components on which *milkER* was based.
3. There were a number of bugs with the BioSQL/BioPerl software and as there is only a small amount of user-oriented development on these projects, this situation was not improved.

Given these drawbacks it became clear that an alternative solution had to be found and the InterMine system was selected to replace *milkER*. The project was consequently renamed *MilkMine*.

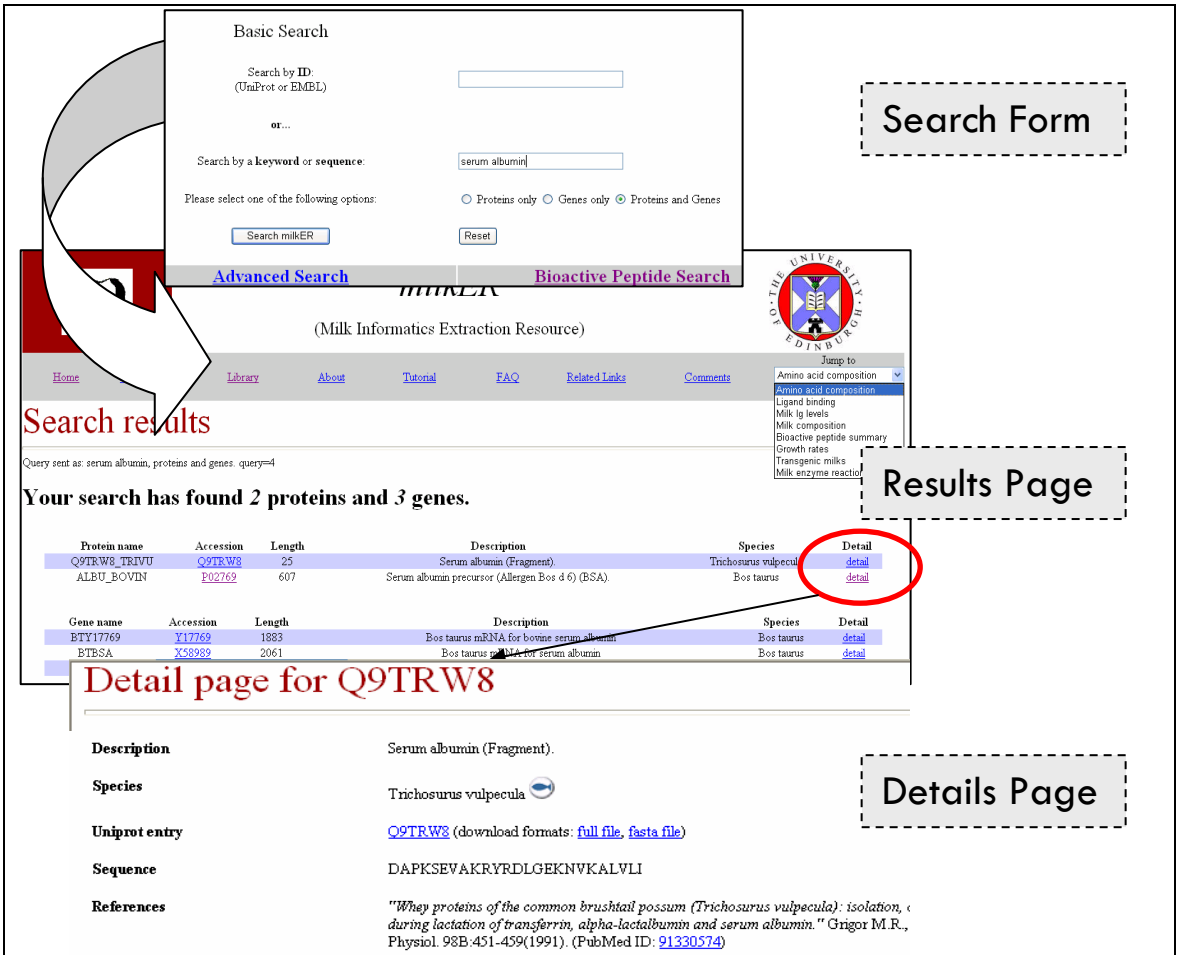


Figure 6.20: The *milkER* database is searchable through a web interface and allows searches for milk proteins, genes, bioactive peptides or sentences extracted from scientific literature. A details page for each protein entry gives additional information such as allergenic information, ligand binding data or links to crystal structures (if available) in the Protein Data Bank.

6.2 InterMine generic system

The standard InterMine model is contained in an XML file which defines all of the possible objects including many biological entities such as proteins, genes and exons. The model can also be extended to allow the user to integrate their own data with the standard data sources. The relational database schema and the web application (webapp) are generated automatically from this model by InterMine scripts.

The power of InterMine comes from its ability to integrate data as they are loaded into the database. For example if a protein entry from UniProt is loaded with the UniProtID, P02754, and a crystal structure from the PDB is loaded with the external linkout to P02754; InterMine will connect the sequence data to the structure data (see Figure 6.21). This data integration provides a powerful platform on which to run complex and diverse queries on the otherwise disparate data. For example, using the illustration in Figure 6.21, a query could be constructed to search for any crystal structures of proteins that are mentioned within a given publication from PubMed.

A. UniProt (P02754)
Beta-Lactoglobulin

B. PDB (1GX9)
Beta-Lactoglobulin with retinoic acid

C. Peptide
Beta-lactotensin (HIRL)
[Beta-lactoglobulin f146-149]

D. UMLS (C0005248)
Beta-lactoglobulin

E. Allergen
Beta-lactoglobulin protein

F. PubMed (15237965)
"*Reduced immunogenicity of beta-lactoglobulin by conjugation with acidic oligosaccharides.*"

G. Secondary integration of data

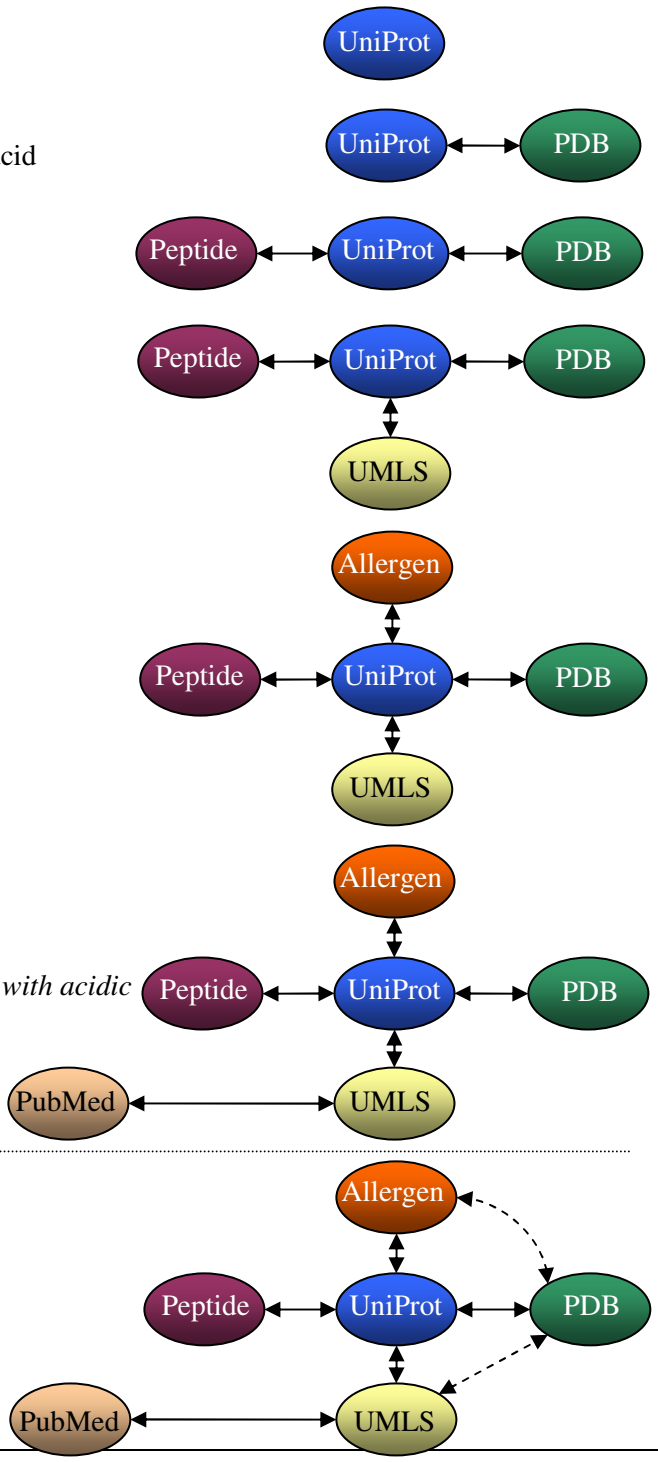


Figure 6.21: As data sources are loaded into the *MilkMine* database, where possible the incoming data are integrated with existing data. Once multiple sources have been loaded, secondary links can also be made between the data (Step G, dashed arrows).

6.3 MilkMine integrated database

The InterMine package was used as the basis to create the *MilkMine* database by following eight major stages (also shown in Figure 6.22):

1. Installation of the InterMine code (and other prerequisite programs *e.g.* Java).
2. Identification of data required for the database.
3. Extension and modification of the standard InterMine model to include *MilkMine* custom data sources.
4. Configuration of the data integration parameters (*e.g.* define which data source should take precedence where there is a conflict between two source databases).
5. Data retrieval from standard external sources (*e.g.*UniProt) and loaded into the MilkMine database.
6. Data retrieval and processing to create custom data.
7. Configuration of the web interface as appropriate (*e.g.* inclusion of aspect pages, title of website *etc*).
8. Creation of template queries as appropriate to allow searching of the database.

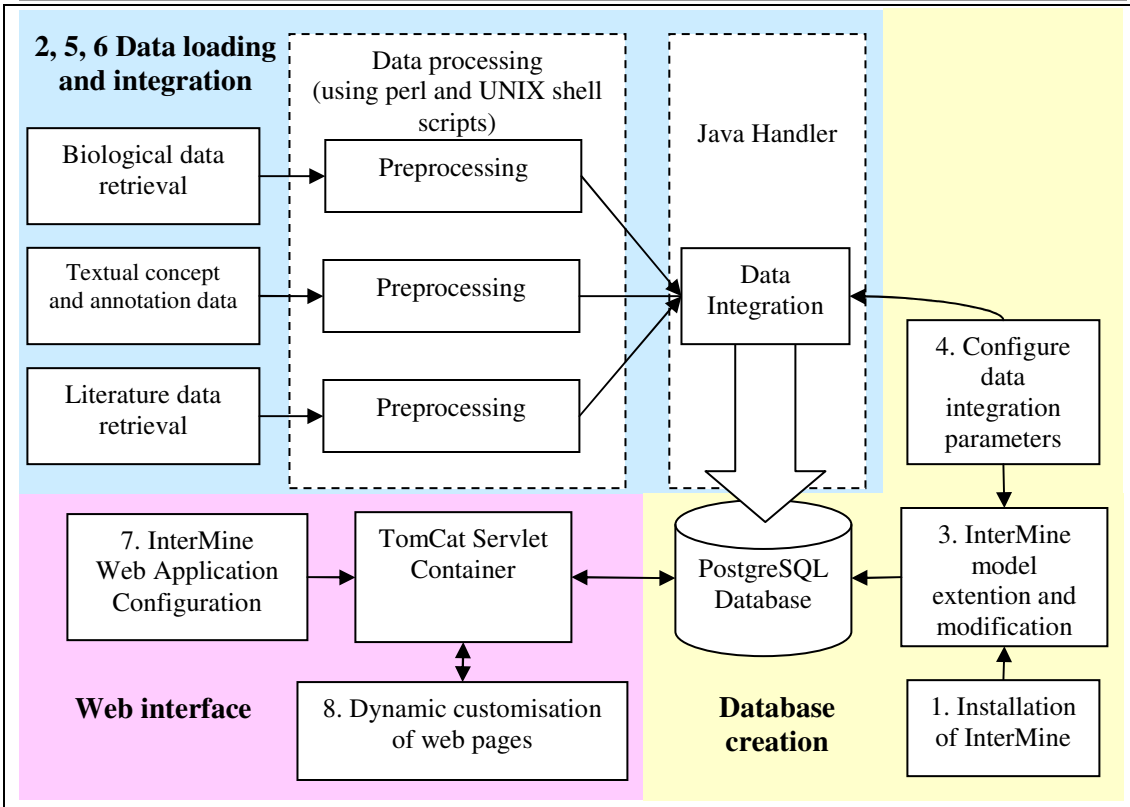


Figure 6.22: Overview of *MilkMine* creation process. The InterMine package was installed, customised and extended. Data was then retrieved from external sources and loaded into the database. The webapp is configured and created. Once the webapp is running common queries and sets of data can be created.

6.3.1 MilkMine data sources

The generic InterMine system is already compatible with a number of common biological data sources such as UniProt, the Gene Ontology and EnsEMBL. However, a number of data sources that would be useful to milk protein scientists or for a literature-mining application were not included in the standard model and had to be added by the author (see Table 6.14).

Table 6.14: Summary of data types included in the *MilkMine* database from a variety of sources.

***MilkMine* Data Source**	**Description**	**Retrieval method**	**Collation method**	***MilkMine* additions required to the InterMine standard model**
Peptide or Protein information				
Milk peptides	Bioactive milk peptide data from scientific literature.	Manual retrieval	Perl script	Yes
Milk proteins	Milk protein data with annotation.	Manual retrieval	Perl script	Yes
Allergens	Milk protein allergen data from scientific literature.	Manual retrieval	Perl script	Yes
Protein structures	Milk protein structural data from the Protein Data Bank (PDB).	Retrieval with Perl script	Perl script	Yes
UniProt	Protein sequence data and annotation..	Retrieval with Perl script	InterMine standard source	No
UniProt-keywords	UniProt specific keyword descriptors.	Retrieval with Perl script	InterMine standard source	No
InterPro	Protein feature annotation.	Retrieval with Perl script	Perl script	Yes
Psi-IntAct	Protein interaction data.	Retrieval with Perl script	InterMine standard source	No
Psi-mi-ontology	IntAct specific keyword descriptors.	Retrieval with Perl script	InterMine standard source	No
Gene information				
EnsEMBL (human, cow, mouse, rat)	Gene annotation.	Retrieval with Perl script	Perl script	Yes
Entrez-organism	Organism data.	Retrieval with Perl script	InterMine standard source	Yes
Terminological information				
MeSH	Medical Subject Headings (including population and age group categories).	Manual retrieval	Perl script	Yes

UMLS	UMLS concept and semantic type data.	Manual retrieval	Perl script	Yes
Gene Ontology (GO)	Gene Ontology data.	Retrieval with Perl script	InterMine standard source	No
GO annotation	Gene Ontology annotation to UniProt entries.	Retrieval with Perl script	InterMine standard source	No
Literature information				
Literature relations (data)	Maps UMLS concepts to literature relations and hypotheses identified by the *MilkMine* text-mining algorithm.	Retrieval with Perl script	Perl script	Yes
Literature relations (sentences)	Maps literature relations and hypotheses identified by the *MilkMine* text-mining algorithm to sentences in the scientific literature.	Retrieval with Perl script	Perl script	Yes
Literature relations (publications)	Maps literature relations and hypotheses identified by the *MilkMine* text-mining algorithm to publications in the scientific literature.	Retrieval with Perl script	Perl script	Yes
Sentences	Actual sentences from the scientific literature, linked to their respective citation.	Retrieval with Perl script	Perl script	Yes
Load publications	Loads citation information for any publication referenced by the other data sources (retrieves information from local storage).	Retrieval with Perl script	Perl script	Yes
Data sources	Information on the data sources used in *MilkMine*.	Manual retrieval	InterMine standard source	No
Journals	Journal data.	Retrieval with Perl script	Perl script	No
Inter-source mapping				
UniProt2UMLS	Maps UniProt identifiers to the appropriate concepts in the UMLS.	Protein names were tagged with UMLS concepts using	Perl script	Yes

		MMTx.		
Protein2Publication	Maps protein entries in *MilkMine* to publications in *MilkMine* using the UMLS concept of the protein.	Protein names were tagged with UMLS concepts using MMTx.	Perl script	Yes
Peptide2Publication	Maps peptide entries in *MilkMine* to publications in *MilkMine* using the UMLS concept of the peptide.	Peptide names were tagged with UMLS concepts using MMTx.	Perl script	Yes
InterPro2GO	Maps InterPro terms to their GO term equivalents.	Retrieval with Perl script	Perl script	Yes
UniProt2PDB	Maps UniProt identifiers (of milk proteins) to their PDB identifiers.	Retrieval of mapping with Perl script	Perl script	Yes
NCBIgene2PubMed	Maps NCBI gene number to PubMed identifier of the referencing publication.	Retrieval with Perl script	Perl script	No
EnsEMBL-entrez mapping	Maps EnsEMBL identifiers to their entrez equivalent.	Retrieval of mapping with Perl script	Perl script	Yes
Sentences2UMLS	Includes tagging information of UMLS terms to sentences extracted from the scientific literature using the MMTx program (see Chapter 4 & 5).	Retrieval of mapping with Perl script	Perl script	Yes
PubMed2MilkMine	Maps the PubMed identifier to the generic *MilkMine* identifier. This mapping allows the use of other literature databases as sources of peer-reviewed publications.	Retrieval of mapping with Perl script	Perl script	Yes
MeSH2publications	Includes MeSH annotation to publications stored in *MilkMine.*	Retrieval with Perl script	Perl script	Yes

Table 6.15: Data included with each data source used in the *MilkMine* database.

***MilkMine* Data Source**	**Name**	**Synonyms**	**UniProt Id**	**Gene Ontology ID**	**UMLS ID**	**Description**	**Milk group and sub-group**	**Organism ID**	**Publication ID**	**Sequence**	**Other data types**
Peptide or Protein information											
Milk peptides	•	•	•	•	•		•	•	•	•	Peptide ID; Parent protein fragment (*e.g.* Beta-lactoglobulin f(120-134)).
Milk proteins			•		•		•				Protein data bank (PDB) ID; structure resolution; structural co-ordinates.
Allergens	•	•						•	•		
Protein structures	•					•		•			
UniProt	•	•	•			•		•	•	•	
UniProt-keywords	•			•		•					
InterPro	•	•				•					InterPro ID; protein feature type (*e.g.* binding site).
Psi-IntAct	•					•		•	•		Psi ID
Psi-mi-ontology	•	•				•					Psi ID
Gene information											
EnsEMBL (human, cow, mouse, rat)	•	•	•					•			Gene length, type and chromosomal location; EnsEMBL ID.
Entrez-organism	•	•						•			NCBI taxon ID
Terminological information											
MeSH	•	•				•		•			MeSH ID; Population groups; Age groups.

UMLS	•	•			•	•					Source vocabularies; Semantic types.
Gene Ontology (GO)	•			•							Term type (*e.g.* Biological process)
GO annotation	•		•	•				•			InterPro ID
Literature information											
Literature relations (data)					•						Literature relation ID; Hypothesis ID; type of interaction (*e.g.* Regulatory interaction)
Literature relations (sentences)					•						Literature relation ID; Sentence ID.
Literature relations (publications)					•						MilkMine (publication) ID
Sentences publications											MilkMine (publication) ID
Load publications									•		
Data sources	•					•					
Journals	•	•									Language; start year; PubMed journal ID
Inter-source mapping											
Protein2Publication			•						•		
Peptide2Publication									•		Peptide ID
InterPro2GO				•							InterPro ID
UniProt2UMLS			•		•						
UniProt2PDB			•								PDB ID
NCBIgene2PubMed								•	•		NCBI gene number.
EnsEMBL-entrez mapping											EnsEMBL ID; NCBI gene number; Rat Genome database ID; Mouse Genome Database ID.
Sentences2UMLS					•						Sentence ID
PubMed2MilkMine									•		MilkMine (publication) ID

MeSH2publications								•	•		MeSH ID

6.3.1.1 Milk protein data

One of the first objectives of the *MilkMine* project was to identify a list of milk proteins. While there is some work which tries to standardise nomenclature of milk proteins (Mather, 2000, Farrell et al., 2004) there is no definitive list of milk proteins (Ward and German, 2004). Therefore a set of standards had to be derived to identify proteins which are of interest. This task has proven to be problematic since the very concept of a milk protein can be defined in several ways:

1. proteins which are present in milk
2. proteins which are overexpressed in the mammary gland during lactation
3. proteins which are involved in lactational processes (including mammary differentiation and involution)
4. proteins expressed in the mammary gland which exhibit a specific signal peptide sequence, directing them for export from the cell.

Each of these definitions of a "milk protein" can be further broken down, for example definition 1 can be sub-divided to:

1. proteins which are present in milk
    a. proteins present at various stages of lactation: early lactation (colostrum), main lactation (mature milk) or late lactation (during involution). This is particularly applicable to marsupial mammals which display a huge variation in their milk composition over the full lactational cycle, largely due to the underdevelopment of their young at birth.
    b. proteins present in healthy or diseased mothers.
    c. proteins present in milk but excluding main blood serum proteins (for example serum albumin) which are present due to leakage into the luminal space of the mammary gland.

Thus these definitions are somewhat arbitrary. Definition 1 was selected as the basis for *MilkMine* as a systematic and repeatable way of retrieving milk protein data.

6.3.1.1.1 Identification of 'milk proteins'

Milk proteins were identified using 3 techniques:

1. UniProt keyword searching

2. Query expansion using UMLS concepts
3. Literature review of key milk nomenclature papers

Identification and retrieval of milk protein data was largely based on keyword searches through the UniProt interface. A system of searches was devised to capture milk proteins from UniProt with good precision, see Table 6.16. This system reduced the protein entries contained in UniProt (version 5) from 2,091,010 total proteins (192,081 mammalian) to 898 milk proteins, representing 80 species.

Table 6.16: Search results for milk proteins performed of the UniProt interface.

Search term*	Search field	Milk vs non-milk protein entries	Precision	Cumulative total
Colostrum	Full text	15/15	1.0	15
Milk	Keyword	137/137	1.0	149
Milk	Comment	46/46	1.0	175
Lactoglobulin	Full text	47/47	1.0	198
E.C. number	Full text	669/669	1.0 (assumed that E.C. numbers were correctly annotated)	788
Ig[A/D/E/G/M] NOT receptor	Full text	115/134	0.86	898
Whey	Full text	47/61	0.77	898
Mammary	Comment	3/11	0.27	898
			TOTAL	**898 proteins**

*All searches were performed with a filter for mammalian proteins only since only mammals produce milk.

A number of problems with using UniProt searches as the only way of defining which proteins are milk proteins, including:

- full text searches can result in false positives where milk is mentioned in a reference title but does not actually refer to the protein or milk mentioned in the gene or protein name.
- comments that refer to milk proteins, particularly whey acidic acid protein which has a common structural domain by the same name.
- using the tissue field of 'mammary' often included breast cancer protein entries.
- a protein *may* be secreted in milk but this location not be reflected in the reference list of the UniProt entry *e.g.* the milk protein osteopontin.

6.3.1.1.2 Retrieval of milk proteins

A Perl script was written to search UniProt and retrieve milk proteins. This means that we can perform semi-automated curation of the database. A more thorough and all-encompassing search is not possible from an automated script since the precision drops and entries must be verified manually.

The 898 UniProt entries were retrieved and then manually verified as being milk proteins. The database cross-references were extracted and analysed. Using the EMBL identifiers from the UniProt records related gene sequences were retrieved from the EMBL database using the EBIs Sequence Retrieval Service (SRS).

6.3.1.1.3 Milk protein query expansion

One of the main problems with UniProt searching is that the functional annotation of milk proteins is incomplete, for instance proteins are often annotated for their function in other cell types rather than as they are in milk, therefore we cannot rely solely on identifying milk proteins through UniProt searches.

In an attempt to complement this incomplete data, the UMLS was used to identify synonyms and symbols of milk proteins. MMTx was used to concept tag the list of milk proteins and also the entire UniProt dataset. By comparing the milk protein concept list against the full UniProt list, further milk protein entries were identified (see Figure 6.23). A small number of manually collated stop concepts were removed from the tagged protein dataset due to incorrect tagging or because they were too general (see Table 6.17).

Table 6.17: Certain stop concepts were identified from the tagged UniProt protein set and were removed from subsequent analysis.

Stop concepts	UMLS Concept ID
B-Protein (TYRP1 protein, human)	C1455498
Protein Precursors	C0033665
Peptides	C0030956
Isoforms (Protein Isoforms)	C0597298
Proteins	C0033684
Gene products	C0751455
Protein Homolog (Homologous Protein)	C1512488

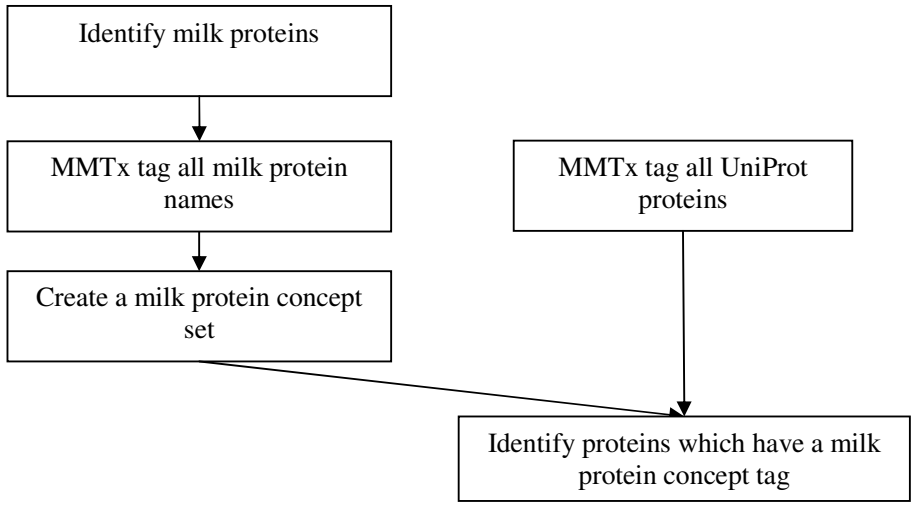


Figure 6.23: Milk protein query expansion using the UMLS concept variants (*e.g.* synonyms and symbols) and MMTx.

6.3.1.1.4 Literature review

A milk protein literature review of key milk protein nomenclature papers (Mather, 2000, Farrell et al., 2004, Fox, 2003b, Farkye, 2003, Fox, 2003a) was also carried out to identify any milk proteins which were not captured by the previous methods.

6.3.1.1.5 Classification of milk proteins

Having generated a list of milk protein entries in UniProt, they were categorised into groups and sub-groups as:

- Casein [Alpha-S1, Alpha-S2, Beta, Kappa] or
- Whey [Beta-lactoglobulin, Immunoglobulin, Enzyme, Lactalbumin, Lactoferrin, Other].

This classification enables the user to create queries that are specific to a particular milk protein, or are more general for a category of milk proteins (*e.g.* the caseins).

6.3.1.2 Protein data

To enable the comparison of milk proteins against non-milk proteins it was essential that these non-milk protein entries were also included in the *MilkMine* database. The entire UniProt database was downloaded from the EBI server as four files (uniprot_sprot.xml, uniprot_trembl.xml, keydef.xml and uniprot.xsd) using an automated script. Protein entries

belonging to a list of 22 milk-producing species (species with at least 5 milk protein entries in UniProt) were loaded into the database, recording information such as protein features, sequence, molecular weight and synonyms.

The UMLS concept to protein mappings from Section 6.3.1.1.3 were also loaded, thus allowing integration of UniProt entries with sentences from the scientific literature.

6.3.1.3 Milk bioactive peptide data

Bioactive milk peptides are a growing area of interest, however as there was no defined source for this data at time of development, milk peptide data was collected from a manual review of the literature. This was aided by a simple peptide extraction Perl script which was written to match standard amino acid sequences (single letter and tri letter notations) and was applied to the milk literature corpus identified in Section 4.1.

Milk peptides were mapped to their UMLS concepts as with the proteins in Section 6.3.1.1.3 .

6.3.1.4 Protein annotation data

InterPro data is collected automatically by a Perl script for key milk-producing species. A Perl script was written by the author to parse the data into *MilkMine* XML format.

6.3.1.5 Protein structure data

Protein structure data are obtained from the Protein Data Bank (PDB) and are integrated with UniProt entries within *MilkMine*. A Perl script was written by the author to retrieve structural models from the PDB and parse them into InterMine format. Since the structural files are often very large only structures for milk proteins are downloaded. This includes key sets of structures for example a set of Beta-lactoglobulin structures from various species and with various ligands. Also included are some milk peptide structures that are available (either as structures in their own right or as ligands). However, the caseins are a family of natively unfolded proteins and therefore are only some theoretical models of their structure (Kumosinski et al., 1993). These are widely disputed (Livney et al., 2004) and therefore have not been included in *MilkMine*.

6.3.1.6 Protein interaction data

Protein interaction data was taken from the IntAct database run by the EBI. Any UniProt accessions within this set are linked to UniProt protein entries for the 22 milk-producing organisms.

6.3.1.7 Genomic data

Genomic annotation information from EnsEMBL was included in *MilkMine* for human, cow, mouse and rat. An analysis was performed to identify the best identifier key to integrate the genomic data with protein data in *MilkMine*. Actual sequence data was not included simply to keep the size of the database to a manageable level; however, every gene can be linked to its respective entry in EnsEMBL.

6.3.2 Textual data

Citations were retrieved automatically from Medline using the NCBI's e-utils tools. An initial search was made using key MeSH headings to create a milk-related literature corpus, giving over 95,000 citations (see Section 4.1). Abundant concepts from this set were themselves used as search terms to grow the literature corpus outwards and thus cover a wide variety of biological literature. The literature-mining work in *MilkMine* has been focussed on the physiological production and metabolism of milk proteins. Technological processing of milk and milk proteins has not been covered although this could be included in further development of the database.

Sentences are extracted from downloaded citations (using MedPost tagger to identify sentence boundaries), are assigned a unique ID and are concept tagged using MMTx (~20,000 articles/day can be processed in this way

Citation data, sentences and MeSH terms are stored locally for fast analysis. Any future literature search results are checked against this local store to avoid duplicating the literature download effort and therefore reducing load on the PubMed server.

Although Medline is the standard literature source in the InterMine system, the *MilkMine* database has been expanded to make it compatible with Web of Knowledge, Chemical Abstracts (CAB) and OVID file formats. Often milk protein or peptide information has come from a reference outwith these literature sources and therefore a further 'other' format is available. External literature database IDs are linked to a central *MilkMine*ID.

6.3.2.1 Co-occurrence relation extraction data

A key aim of the *MilkMine* project was to incorporate literature-mining data with standard biological data. Therefore the analysis performed in Chapter 5 was valuable for inclusion in *MilkMine*. Co-occurrence literature relationships are stored as two interactor concepts with a relationship strength value.

6.3.2.2 Terminological data

A number of popular terminological resources were included in the *MilkMine* database.

6.3.2.2.1 Medical Subject Headings (MeSH terms)

The entire MeSH set of terms was downloaded from the NCBI and parsed into *MilkMine* format for inclusion in the database. This allows improved semantic searching of the database, for example sentences in the database can be assigned to a particular organism depending on the MeSH terms assigned to the article.

6.3.2.2.2 Unified Medical Language System (UMLS) metathesaurus

The UMLS 2006AB version was used for *MilkMine* including definitions for many of the terms. MeSH terms are mapped to their UMLS concepts. This allows the integration of citations with the UMLS through any MeSH annotation. Semantic filters were also created, for example 'Biological Entity [Gene; Protein…]' so that data in *MilkMine* could be searched in coarse or fine granularity.

6.3.2.2.3 Gene annotation data

The Gene Ontology and Gene Ontology Annotation (GOA) data are included in *MilkMine*. The gene ontology is simply the record of GO terms and their description and structure. The GOA data source links these GO terms to other biological entities, for example InterPro or

UniProt entries. The entire set of Gene Ontology terms is entered into *MilkMine* whereas GOA is added for the key milk species only.

6.3.2.3 MilkMine web interface

The *MilkMine* interface is designed to be as easy-to-use as possible while still providing complex and flexible functionality, for example "Search sentences by amino acid sequence". This query allows the user to work with completely raw amino acid sequence data to retrieve knowledge without knowing protein names, synonyms or even identified domains.

Data in *MilkMine* can be accessed in a number of ways at varying levels of user involvement, for example datasets are available for simple download, pre-completed template queries can be used to search the database or users can create their own database queries from scratch. The wide range of features available include (the web interface is covered in detail in the *MilkMine* tutorial, see Appendix I):

- A point and click query builder can be used to modify existing template queries or create user specific queries.
- Customisable search result tables can be ordered or altered as the user requires.
- Previous queries can be saved.
- Query result tables can be combined, intersected or subtracted or even used in subsequent queries.
- Sets of data can be stored in 'bags' by the user to use in subsequent queries, to download or to view at a later date.
- An on-line protein structural viewer, JMol can be used to view milk protein structures within *MilkMine*.
- Data can be output in a variety of formats (*e.g.* FASTA format for subsequent use in other bioinformatic programs such as BLAST *etc*).
- Print and video tutorials as well as help documentation are available on-line.

The *MilkMine* webapp has been tested on a variety of web browsers and versions. The most popular web browsers for using *MilkMine* are Firefox Mozilla 2.0 (58%) and Internet Explorer 7.0 (22%). Both of these browsers are compatible with the *MilkMine* interface and have only minor aesthetic differences in web page representations. One main requirement for the user is

that their web browser is java enabled since this is required to run the embedded JMol protein structural viewer.

6.4 MilkMine database updates

The update of *MilkMine* is an important aspect of the project since biologists need to have up-to-date data. Ideally this will be good enough for auto curation but should also be flexible for future terminology changes, for example deprecation of the keyword "milk" in UniProt records.

6.4.1 InterMine source code and database schema updates

The InterMine system which underlies *MilkMine* is being actively developed and therefore new releases are continually produced, *MilkMine* is currently uses InterMine version 8.2. The InterMine code was downloaded from the InterMine website and any bug fixes or development can be updated on *MilkMine*.

6.4.2 Database content updates

The data retrieval and update process is automated as much as possible to reduce overheads and enable more regular updates. Automated scripts retrieve up-to-date data (see Table 6.14) which keep a track of dataset versions. This means that updates can be done much more easily and therefore more frequently. Each new database build takes approximately 2.5 days to complete.

Data is loaded and integrated in a construction version of the *MilkMine* database and is then copied to a production database which can be accessed through the web application (http://www.bioinformatics.ed.ac.uk/milkmine). This ensures that changes and updates can be tested and corrected with minimal disruption to the working database. The database can be dumped at various points during the data load process therefore any load errors do not mean that you have to start again from scratch. Although this reduced the development time to create the *MilkMine* database, the installation and development took a substantial amount of time.

Literature updates

As areas of interest are studied, literature searches provide new citations which are retrieved, tagged, parsed and stored (max 10000 articles per search). As the local database of citations is expanded, coverage of areas of literature relevant to milk protein scientists increases.

Relationship updates

Once the literature has been updated, the biological relationships also need to be updated. A list of the concepts tagged to the newly updated literature is produced and any relationships involving these concepts are re-evaluated.

6.4.3 Implementation and performance

The *MilkMine* database is hosted on the Edinburgh Centre of Bionformatics (ECB) server cluster, a Dell Power edge 2850 (Dual Intel Xeon processor 3.4Ghz, 4Gb RAM server). The Perl benchmark utility was used to time key parts of the text-mining algorithm and MilkMine scripts to identify computational bottlenecks which could be optimised as required.

The size of many of the datasets loaded into *MilkMine* are extremely large, and therefore some optimisations were made on loading the data to avoid running out of memory or taking a prohibitively long time to load. During development small datasets were loaded and tested before loading the whole set of data.

Chapter 7 - Discussion and conclusions

7.1 Automatic extraction of domain related terminology

The complete and precise identification of terminology within the scientific literature is a key step in the success of any literature-mining algorithm. For the *MilkMine* project, the MMTx program was evaluated and found to be proficient in performing biological Named Entity Recognition on MeSH terms (Recall=0.93 and Precision=0.82, see Section 3.3). The errors found in the MMTx performance analysis were largely the result of incorrect parsing of the terms into phrase chunks; particularly those with longer and more complex structures.

A small number of common phrases were found to be repeatedly overtagged by MMTx *i.e.* they received a UMLS concept tag in error. These were filtered out by the author using an automated method (see Section 3.4) based upon the expected document frequency of the UMLS concept using a search in the Medline database. Any concept that had a much higher document frequency in the MMTx tagged sentence set when compared to the document frequency in Medline was removed. Although this approach removed many of the commonly mistaken taggings, it did not completely remove this source of error from the *MilkMine* system.

A further simple improvement to the performance of MMTx was to force the program to properly recognise the scientific notation of organisms. Given the format '*S. cerevisiae'* MMTx would split the name into two separate phrase chunks. Replacing the shortened form of the name with the full name increased the Recall and Precision to 1.0 (where the organism was represented in the UMLS).

One of the main problems with the MMTx performance analysis (see Section 3.3) is that it was based on a list of MeSH terms alone rather than on free text sentences. Therefore it may have been preferable to have tested the program on a corpus of free text sentences. Although this would have been much more labour intensive to achieve it would have provided more realistic data on the performance of MMTx under the conditions that it was used within *MilkMine* (*i.e.* tagging literature sentences). A further evaluation method that could have been employed would have been to compare the output of MMTx against the MeSH terms assignations given by

genuine MeSH indexers to determine if MMTx produced a similar performance to human annotators.

The second main problem experienced when trying to use MMTx to tag UMLS concepts to text was that the concept *had* to be represented in the UMLS Metathesaurus. As the Metathesaurus is designed to be applicable to a wide variety of uses and subject areas, it often lacks terminology to a specific domain or sub-domain. Therefore, it was important to be able to accurately identify terminology that would be of interest to milk protein and peptide researchers.

Selection of milk related literature

Milk related literature was first selected using MeSH term searches in Medline. MeSH terms can easily provide a corpus with high precision and recall - assuming that there is a relevant MeSH term available and has been in existence long enough to build up a corpus. Four MeSH terms (Milk; Milk Protein; Lactation and Colostrum) produced a corpus of over 95,000 publications (see Table 4.9) however, the sub-literatures were substantially distinct (86%). This result was somewhat surprising given the expected cohesion of the four sub-literatures and shows that while each of the four sub-domains could be thought of as being 'milk-related', they display a diversity which may prove productive in the application of literature-mining techniques.

Literature Gradient Technique (LGT)

The Literature Gradient Technique was used to identify terminology which was significantly more prevalent in the milk related corpus when compared to the document frequency of the terminology in general biological literature. Ranking the journals simply by number of Milk[MH] publications or by milk related value (MRV), gave a slightly different but more meaningful order to the top journals. For example the *American Journal of Clinical Nutrition* ranks second by number of Milk[MH] publications is the although it only has an MRV of 5.20%. The *Journal of Dairy Research* ranks higher by MRV (57.85%) than the *Journal of Dairy Science* (39.87%) despite having published significantly fewer Milk[MH] articles. Therefore using MRV to define the probability of retrieving milk related publications from a

given journal is a more accurate method to create corpora 3; 2 and 1 than using a straight frequency metric.

The rules used in the literature gradient technique to define the categories of milk related literature were based on two variables, number of journals and subject specificity (see Table 4.10). This situation was not ideal as the two variables should have been isolated and examined independently to determine the effect of each parameter on the literature gradient. However the author tried to address these issues by normalising the term frequency to the total number of terms in each corpus. 8,500 publications were used for each corpus and each gradient level to provide good coverage of terminology.

From the two variable gradient, it was interesting to note that both variables had a bearing over the redundancy of terminology within their respective corpora (see Figure 4.14). The analysis shows that the variation in terminology within a domain is greater than within a specific journal. This should therefore be an important consideration when evaluating literature relationship sets such that the strength of a particular relationship will depend upon both the number and the similarity of the journals from which publications are included in the analysis.

The author attempted to address any relationship distribution skew by introducing a filter which downweighted the score of high, narrow sentence co-occurrence relationships (*i.e.* relationships which had a very high frequency of occurrences but within a very small number of publications).

Identification of milk-related uni-gram terminology

The Termine application does not identify uni-grams during terminological analysis therefore this had to be completed using the Literature Gradient Technique method. Using a cut-off limit of 10 x more frequent in milk-related literature over general literature was successful in producing a good array of milk-related uni-grams (single word terms), however the method relied upon the term appearing in both milk-related and general literature. Therefore, some common milk-related terms (for example ‘breast-feeding’) were not identified by this method purely due to the fact that the term did not appear in the general term list. To catch more of these terms, category 4 terminology (highly milk related) was compared against the general corpus

(category 0). Although there would be more similarity between categories 4 and 0 than between 5 and 0, there was still enough significant difference to differentiate important milk terms.

A simpler method to identify milk-related uni-grams may have been to calculate the TF*IDF value for each concept tagged in the milk-related literature corpus identified in Section 4.1. This would then be repeated for each of the concepts over the whole Medline database. Thus an accurate TF*IDF value could be obtained for each concept within milk-related literature *and* within general biological literature making a complete comparison possible (*i.e.* it would negate the problem of not having the milk term within the general list).

Identification of milk-related n-gram terminology using Termine

The Termine application used a much more complex set of metrics to identify n-gram (multi-word) terms than was seen with the uni-gram analysis. An evaluation of Termine performance over citation size was completed to ensure that the author was accurately identifying terminology that was significantly milk-related.

Although the absolute C-value[1] was found to increase linearly with corpus size (see Figure 4.15), the relative C-value (see Figure 4.16) levelled out above 750 citations. This agrees with Termine's own documentation which states that it needs a "fair amount of text to produce reasonable termhood scores, as they rely on the term appearing multiple times within the corpus". The maximum number of citations allowed by the Termine web interface was found to be around 1,250 citations. As this was above the 750 citations required to achieve a consistent relative C-value, 1250 citations were used to perform the Termine analysis on milk-related literature.

Combined results from term extraction

As described in Section 4.2.3, the top 500 terms for Milk[MH], Milk Protein[MH], Lactation[MH] and Colostrum[MH] literature corpora were identified leading to 3,611 distinct

[1] The C-value is the score given to a candidate term within a corpus to indicate the likelihood that the term is significant within that corpus. The higher the C-value, the more significant the term is.

terms in total. 512 (14%) of these terms were not represented in the 2006AB UMLS Metathesaurus; 750 (21%) were new variants or spellings of existing concepts and 2,394 (66%) were already present in the 2006AB_custom UMLS dataset.

It was encouraging that a significant proportion of the concepts identified were already represented in the UMLS Metathesaurus, demonstrating the scope and applicability of the UMLS to biological projects. However, a substantial number of concepts (512) were not found at all or were identified as new variants (750) which would not be recognised by MMTx or would at least receive a lower score. Therefore the complementation of the UMLS with milk related terminology resulted in a 35% increase of the milk-related terminology contained within the UMLS as identified in Section 4.2.3.

7.2 Complementation of the UMLS for the MMTx program

Using UMLS concepts to perform NER in a text-mining system is beneficial as synonyms and lexical variants are mapped to the same concept and describe its semantic type. However, if there is no UMLS concept available then obviously there will be zero chance of identifying any literature relations or hypotheses for that concept. Therefore, as many key milk related concepts identified in Section 4.2.3, such as 'mammary gland secretory epithelium', 'skin-to-skin contact' and 'Thyrotrophin-releasing hormome' were not found within the 2006AB dataset, it was important to attempt to create a complemented milk UMLS subset so that key milk related concepts would be represented.

One of the main issues with using the UMLS Metathesaurus as a basis to perform NER with the MMTx program is that it does not contain a definitive list of protein or gene names. There are 97,136 protein concepts within the 2006AB UMLS dataset and are therefore used to tag literature sentences for the *MilkMine* text-mining algorithm. However many of these protein concepts are more general than specific, for example the UMLS concept 'alpha-caseins' existed but there was no differentiation between alpha-S1 casein and alpha-S2 casein proteins.

Clearly for the *MilkMine* algorithm to be optimum it requires accurate identification of specific proteins and thus be able to distinguish between the functionality of a protein as distinct from the class of protein. The authors attempted to resolve this by creating a separate protein name

gazetteer (list) which could be used alongside the MMTx tagging. A pattern-matching Perl script was written to match this list against the literature however, this was extremely slow with poor precision and recall, and therefore was not pursued. Milk related protein names identified in Section 6.3.1.1 were included in the UMLS complementation process but non-milk related proteins were not.

Potentially a third-party NER software, such as ABNER could be used to improve the protein and gene NER within the literature sets used in *MilkMine*. These identified proteins could be then be analysed within the literature-relation extraction algorithm as pseudo-UMLS concepts.

It is important that the creation of a new customised and complemented UMLS is as straightforward as possible so that it can be updated and maintained as well as being applied to other domains of scientific literature. This is partly assisted by the provision of files containing the changes made between one UMLS version and the next. Although a new version of the UMLS is released 4 times per year, not all of the source vocabularies are updated on every release. Furthermore, any additions made to the UMLS Metathesaurus by *MilkMine* (this also ensures compliance with the UMLS copyright laws) are given a unique code so that these additions can be directly added to a new release.

7.3 Co-occurrence relation extraction in *MilkMine*

Literature-mining remains an incredibly difficult goal to achieve with a high degree of accuracy due to the inherent variability of natural language. Co-occurrence based relation extraction provides are relatively simple to implement and produce better recall than natural language processing (NLP) methods as they can identify relationships between any two concepts. However this high recall comes with a correspondingly low precision, although there are methods to improve this such as filtering the set of literature relations produced (see Chapter 5). For these reasons, co-occurrence relation extraction algorithms are well suited to exploratory methods.

The UMLS dataset does come bundled with a file of co-occurrences of UMLS concepts in the Medline database. However, this was not used as it did not contain the milk related terminology as identified in Section 4.2.3) concepts. These relationships were also based on MeSH terms

only and therefore represented a small proportion of the UMLS. As MeSH terms are assigned for only the major concepts of the article, this excluded the more incidental and therefore potentially more interesting concepts. The file is updated with every UMLS release (*i.e.* once per quarter), however, there is no flexibility with using this data in comparison to a local text-mining algorithm, for example to analyse literature relationships over specific time periods *e.g.* 1960-1986 as with Swanson's initial Raynauds' Disease – Fish oil connection. For these reasons a local algorithm was developed.

The selection of filters used in the *MilkMine* literature-mining algorithm were relatively straightforward to implement but still greatly reduced the number of literature relations and hypotheses identified. In the current implementation of *MilkMine* containing over 5 million sentences from the scientific literature, only 500,000 literature relations were generated. This reduction was produced largely through the exclusion of concepts from the literature-mining analysis (*i.e.* static filters such as stop concepts or stop semantic types, see Table 5.13). This reduced the load on the perl script which actually performed the analysis. Relation-based filters such as maximum allowed number of relationships from a given concept were also effective in removing highly documented and therefore less interesting relationships.

However, despite the measures used to remove false positive relationships from *MilkMine*, there are a number of drawbacks to relying on entity co-occurrence within text as a method of identifying a relationship between those entities. The rate of false positives is high, particularly over longer or more complex sentences (although these also represent key sources of error for rule and machine based relation extraction methods).

Citation size is important in text-mining as there needs to be enough completeness of data to identify a relationship but too much data is also a problem. Swanson found that the optimal size for hypothesis generation using open or closed discovery was between 100 to 5000 articles (Swanson et al., 2006). Clearly, if a biological concept has a publication collection of over 40,000 articles it is not feasible to include all of the interactions contained within them. Also, a majority of the data within this article set will be redundant or out-of-date and so a maximum limit of 500 articles were set for the literature retrieval step for each term. This compares favourably with other literature-mining systems such as ChiliBot (Chen and Sharp, 2004) with a

default setting of 50 articles per term pair and EBIMed (Rebholz-Schuhmann, 2005) with default setting of 500 articles (see Appendix B for a comparison of literature-mining systems).

The categorisation of literature relations and hypotheses using set lists of interaction verbs is also useful for the user to be able to hone in on information of interest. Rather than reading through all of the literature relations on their chosen concept, they can simply look at certain *types* of interaction (*e.g.* regulatory interaction, see Section 5.3.4).

Implementation issues with the MilkMine co-occurrence relation extraction algorithm

Co-occurrence relation extraction cannot identify the directionality or type of relationship *i.e.* an A → B association is not necessarily the same as a B → A association. Although this could be partially solved by defining the direction of the relationship as being from the query term first mentioned in the sentence to the query term mentioned second (Chen and Sharp, 2004), this is not reliable. The actual number of co-occurring sentences between 'A and B' and 'B and A' will be the same since by definition they both have to occur in the same sentence. However, the significance of A in the B literature and B in the A literature will be different, particularly if the two search sets are very different in size. Performing co-occurrence extraction first followed by rule-based or machine-learning based relation extraction should provide high recall *and* precision of milk protein relationships, with directionality and a reduction of the false positives found by co-occurrence. These analyses are beyond the scope of the current thesis but could be applied as future work to further increase the accuracy of the *MilkMine* algorithm.

MilkMine relies on sentence co-occurrence to determine whether or not a relationship exists. However, there is no record of the position of the biological entities (*i.e.* UMLS concepts) within the actual sentence. This level of detail could potentially allow more complex analysis of the relationship between the co-occurring entities such as calculating the distance between the two entities or if there is a verb that verb defines the relationship. Requires saving a much larger amount of data for the phrase (could do by character position (*e.g.* phrase is from character 45 – 56). In the i-HOP system, biological entities are mapped to and hyperlinked from their exact position in the text *e.g.* "The normal **mammary gland** of a multiparous woman is characterized

by several known differentiation markers such as casein kappa, casein beta, keratin 14, CCAAT/enhancer binding protein beta and delta and adipsin [?]. [2005]".

MilkMine makes use of organism MeSH terms assigned to publications in Medline. Often the organism is not specified in the abstract but is mentioned in the full text of a publication. Therefore literature sentences can be retrieved limited by organism, however the organism classification is not included in the relation extraction or hypothesis generation algorithm *e.g.* a literature relationship, 'Protein A -> Protein C' may not be accurate since the interaction may be species specific for a number of reasons. For example Protein A from cow may have a different structure to the human version and may display a set of physiological interactions in cows but not in humans. While this again reduces the reliability of the set of literature relations generated for *MilkMine*, the identification of the source organism for each protein mentioned in text is beyond the scope of this thesis. Similarly, *MilkMine* used sentence co-occurrence as the basis of relation extraction, as have many other systems such as EBIMed, i-HOP and ChiliBot; but relationships can also be extrapolated across sentence boundaries (known as anaphora resolution). However, Ding *et al* found that the benefit gained of using anaphora resolution was not worth the effort of development (Ding et al., 2002) and therefore this was not included in the current work.

When attempting to create a text-mining system specifically for a particular biological domain, a trade-off must be reached. Obviously the key literature was identified, tagged and analysed but this process was extended outwards to cover more and more literature. There is a problem of identifying literature relations from incomplete sentence sets, particularly where there is a disproportionality between the two sets. For example the literature for the milk peptide beta-casomorphin may be stored entirely in the local database. When hypothesis generation is performed from the concept 'beta-casomorphin', an intermediate concept is identified as 'insulin'. There will be some literature stored in the local database for insulin from previous search expansion; however the entire collection of insulin literature will not be contained in the local database. Therefore when any hypothesis generation is made from insulin, this will be based on only a part of the insulin literature. Therefore, there is some disparity between the amount of coverage being made between the respective literatures. One way around this issue would be to calculate a ratio of how complete the local literature database is for a particular

concept set retrieved is (*i.e.* out of a PubMed search, how many do we have in local store and therefore how many are involved in the relationship analysis?). 100% of all publications in the Medline database for a given UMLS concept will never be achieved as that would involve having the entire Medline database in local store.

One final issue which affects all relation extraction systems is the issue of negation sentences. These reduce the accuracy of systems whereby false positive relationships are accredited where they should be true negatives. Thus the precision of the algorithm is reduced while recall is unaffected. Removing any negation sentences would therefore improve the precision by reducing the number of false positives but in doing so the recall will be compromised. Literature from negation sentences can be removed by pattern matching negation terms *e.g.* 'not', 'but', 'whereas', 'although', 'however' etc.

7.4 Domain specific issues when creating a literature-mining system

As has been described, the *MilkMine* system has followed a generic method such that the processes used to set up the integrated database and text-mining algorithm should be applicable to other domains of research. However, there will be particular issues with any given domain which should be taken into consideration when creating a literature-mining system. One of the key issues for *MilkMine* was to identify proteins which are 'milk related'. The definition of a milk protein used (see Section 6.3.1.1.1) was not intended to be extremely precise but simply to produce as high a recall as possible. As a 'milk protein' can be a rather arbitrary concept (*e.g.* should it represent *only* those proteins found secreted in milk or should it include lactational machinery proteins?) it was impossible to calculate recall figures. Recall and precision figures could be derived however for individual key word searches. As with the identification of milk related literature in Section 4.1, a few key words were found to return the largest number of milk proteins.

In attempting to retrieve more minor and less well documented milk proteins, the precision of keyword searches dropped dramatically. For instance, if a protein was documented as a milk protein for one species this was assumed to be representative for other species. However, this is not always correct, the key example being beta-lactoglobulin, an important and abundant milk protein in bovine and ovine milks but is completely absent in human, equine and murine milks.

In this case this absence is caused by a lack of the beta-lactoglobulin gene and not because the protein is not excreted in milk. It is very difficult and labour intensive to assign accurately a definitive list of milk proteins for any given species. There is also an inherent problem of incomplete data among species. Common milk producing species, particularly the cow have a much larger list of documented and characterised milk proteins and peptides.

Another issue that is more significant for milk protein relation extraction than other protein sets is that several of the major milk proteins (*e.g.* casein, β-lactoglobulin) are commonly used as standard proteins and therefore false positive relations may be produced. However, the use of milk proteins in this context will more likely appear in the materials and methods section of publications rather than in the abstract, therefore this is less of an issue for *MilkMine*.

The size of a domain is also important for relation extraction. This was a problem found with the milk bioactive peptide literature where many peptides are described in only a few publications.

7.5 Application of an integrated database system

When designing a system like *MilkMine*, there is a trade-off which must be balanced between database warehousing and federation. A local copy of data can be organised and manipulated is a much more customised way, however it requires continual updating so as not to become out of date. As with any biological database, it must be updated on a frequent basis. *MilkMine* has been developed with a view to being responsive to:

1. new data (*i.e.* in the raw data which is entered into the database)
2. new data types (for example, protein structural data was added to the *MilkMine* model)
3. data structure changes (*e.g.* the PDB format is due for a change soon).

Other text-mining systems, noteably i-HOP (Hoffmann and Valencia, 2004) and EBIMed (Rebholz-Schuhmann et al., 2006) use local storage for fast retrieval, parsing of citations and implementation of the web interface. *MilkMine* is a similar local store system but contains a much wider variety of data types. The real power of *MilkMine* however is the fact that each dataset is integrated into the entire model, therefore allowing much more complex and deep querying of the data than would otherwise be allowed.

MilkMine was the first external use of the InterMine package, the team was keen to provide technical support for the creation of *MilkMine*. Although developed primarily by computer engineers it is very much aimed at biological users. Through the application of *MilkMine*, many features and improvements have been made to the InterMine structure, design and documentation. While the use of the InterMine system allowed much more complex functionality than could have been achieved through the authors own design, there was still a substantial amount of investment time required.

The use of third party software has advantages and disadvantages. The InterMine system is being developed by a larger team of experienced software engineers and therefore provided much better database design, more advanced features and quality control than the author could have produced himself. However, this also created a reliance on the funding, input and priorities of others, which may not have matched those of *MilkMine*. For the purposes of this project, the two objectives were found to marry up acceptably.

Unfortunately, integrating data has the potential error of inaccurately ascribing information, particularly when the data has come from a hierarchical source. For example, if a mapping is made from a specific to a more general term, there is not necessarily a legitimate mapping in reverse. This will reduce the legitimacy of many of the literature relationships identified by the system and data, and therefore as a consequence the reliability of the entire system is reduced. However, *MilkMine* is intended to be a beneficial aid for researchers and therefore any results obtained should be analysed further to ensure their accuracy.

Another problem that was found in creating the MilkMine data warehouse was economy of scale, particularly with reference to the literature-mining data sources used. Increasing the size of the sentence set used to generate the literature relationship set led to a huge increase in literature relationships and hypotheses generated to be included in the database. However, each relationship is only stored in the database once (with links to the evidence sentences) and therefore relationship redundancy increases with citation set size.

7.6 Biological evaluation of the system

MilkMine was created largely as a prototype system to show that it was possible to integrate standard biological data types with literature-mined data. Many milk protein researchers have registered an interest in the development of *MilkMine* and have provided feedback on the system. However, more evaluation would be required to prove the effectiveness and applicability of the overall system and interface.

Despite the attempt to create a domain related literature-mining and integrated database package, the area of 'milk science' may still have been too ambitious as it covers a huge diversity of science from dairy processing to complex molecular physiology. For example, feedback was received from Prof. Douglas Dalgleish (University of Guelph, Guelph, Ontario) that there was not enough data on dairy chemistry and processing. Although this was due to a deliberate decision to concentrate on physiological aspects of milk proteins and peptides, it demonstrates the abundance of scientific literature. The current implementation of *MilkMine* contains over 5 million sentences and 500,000 literature relationships, however even at that scale some areas of science can only ever have small coverage in *MilkMine.*

The optimal text-mining system would only specify those links which are genuinely and biologically possible. To do so the system needs to recognise all of the following conditions in any given instance: temporal, spatial, physiologically significant, and thermodynamically possible, *in vivo*/vitro, natural or manipulated (*e.g.* genetically engineered). For example, for a direct protein-protein interaction to occur, clearly both proteins need to be in the *same place* at the *same time*. However, even if this was achieved it is difficult to ascertain if a bioactive peptide of interest actually gets into the blood stream in an active form or whether it is broken down within the gut. To enable this is beyond the scope of the current work although future literature-mining systems may be able to cope with this level of granularity.

One potential improvement that could be applied would be to remove any sentences containing '*in vitro*' as these sentences are more likely to contain less probability of physiologically significant interactions. However, it will also reduce recall of true positives such as sentences of the format 'Protein A was shown to interact with protein B *in vivo* but not *in vitro*.".

While clearly it would be ideal to have all of these factors in the system, it may be some time before all of these factors can be included in any given text-mining system. *MilkMine* took the attitude that it would present potential hypotheses whereby the user can directly check the evidence for such an assertion. Therefore, while we do not claim a perfect system, the user can investigate any relationships or hypotheses that interest them.

7.7 Abstract *vs* full text

There is a debate as to the optimum unit of literature to use in text-mining systems. Abstracts have been shown to have an order of magnitude more information than titles alone and have been shown to contain the highest density of key points within a publication (Kostoff et al., 2004). In theory abstracts should be a concise conclusion of the key results of the paper therefore providing a good basis for extracting the key relationships contained in the article. However the full text of a publication contains much more data as well as containing more speculative text (*e.g.* in the discussion section) and lesser known pieces of data than are contained in the abstract. Given that hypothesis generation looks to pull together disparate relationships there may therefore be more knowledge discovery potential with the full text of publications.

However, there were several reasons for restricting the unit of analysis in *MilkMine* to abstracts, the main reason being the issue of scalability. Using co-occurrence relation extraction, an increase in the number of sentences will greatly increase the number of relationships identified. Although there is more repetition of literature relationships within the full text, the number would become crippling. More work is required to refine the generation of candidate literature relationships and hypotheses produced by the *MilkMine* text-mining algorithm before it can be scaled up to full text.

There are also a number of practical issues such as the fact that full-text articles are much less available, require more processing (for example PDF conversion) and display a much higher variation in topology (for example figure legends, tables and non-standard characters *etc*).

To evaluate the improvement in relation extraction and hypothesis generation by using full text rather than abstracts, an example case study could be analysed, comparing the recall and

precision for the extraction of a known hypothesis. This would determine whether there is value to be gained from expanding to full text or whether the processing and storage overheads are too expensive. Although there is the possibility of using abstracts and full text within the analysis, correction methods would need to be in place such that the repetition of a relationship seen in a full text article does not dominate over a relationship extracted from abstracts within the corpus. For this reason it would be suggested that either abstract *or* full text articles should be used for text-mining analysis but not both.

7.8 Impact of *MilkMine* on scientific knowledge

MilkMine joins a growing list of integrated database and literature-mining systems, however, there are several unique points that the project brings to increase scientific knowledge. While many systems provide literature-mining which is linked to external databases, *MilkMine* is the first system to fully integrate biological data types with data derived from literature-mining. *MilkMine* also represents the first implementation of the InterMine architecture by a third party and was therefore instrumental in the testing and implementation phases of that product.

MilkMine has an advantage over many other text-mining systems in that it is accessible to biologists (*e.g.* those who are simply looking for articles or protein sequences *etc*). They can access text-mining data at a very basic level by simply viewing related concepts; clicking through links in the interface or they can get more involved and start studying generated hypotheses.

While the *MilkMine* literature-mining algorithm may not represent new techniques within themselves, the semi-automated generation of domain related terminology is widely applicable and repeatable. By using UMLS concepts and looking at terms which are over-represented in milk literature compared to general literature, the need to manually generate lists of stop words and keywords was removed, thus negating the long and laborious process involved. Therefore, the technique is much more transferable to other domains.

Many text-mining systems only look at protein or gene interactions, however *MilkMine* looks at interactions between many biological concepts such as diseases drugs, processes, proteins and genes. Use of bioinformatic methods in general have become increasingly accepted as research

methods, however, the application of end-user systems has been lagging behind this. Looking at two examples, ARROWSMITH (Smalheiser, 2005) and ChiliBot (Chen and Sharp, 2004) have reasonable interfaces, and have been cited 37 and 29 times respectively. I-HOP (Hoffmann and Valencia, 2004) has been cited 92 times. While *MilkMine* is by definition a tool primarily for milk protein researchers, the applicability of the concept of *MilkMine* has the potential to reach beyond this. *MilkMine* is being used as an example implementation of InterMine and has received interest from researchers of other fields.

7.9 Future trends and directions in this field of research

For literature-mining to become much more accepted as a viable method of research, it must become more accurate and robust. Proper mark-up of key concepts is required in a standardised way, by the author (therefore avoiding misinterpretation) and at point of publication. For example marking up every occurrence of a gene name with a stable link to existing biological databases. This will make text-mining and evaluation of text-mining systems significantly easier, more accurate and therefore more meaningful (Shah et al., 2003). Standardised, agreed and implemented standard formats are essential for interoperability of these systems. Recent projects such as SciXML and SBML are standards which aim to facilitate this move towards an enriched format. Systems utilising text-mining techniques will become more popular as they improve in accuracy, usability and relevance.

The ideal is that text-mining systems will overtake standard literature database interfaces as the first point of reference for researchers, however they must prove their accuracy first. This will require a substantial development phase and text-mining will not become standard until the recall, and more importantly precision of systems, have reached a level approaching human comprehension.

In future, literature search engine output displays will continue to move away from the simple citation view towards a semantically enriched display in which inter- and intra- literature relationships are presented in a cognitively enriched manner, by more intelligent representation of relationships. The level of personalization of these interface systems will improve dramatically, becoming much more customizable (as with *MilkMine* user-definable templates, storage and search query histories) and user-friendly.

7.10 Conclusions

The vast and increasing volume of biological data can make it a struggle for scientists to keep up-to-date with the latest research and as a consequence they may miss significant biological links, particularly those that extend outwith their own area of expertise. *MilkMine* is an attempt to provide a single informatics resource to help milk protein scientists mine this information mountain more effectively, by integrating standard experimental data types with data generated by emerging text-mining techniques.

A method was initially developed to identify milk-related terminology from peer-reviewed biological literature and this was used to complement the Unified Medical Language System (UMLS), a large thesaurus of biological concepts, their variant names and their semantic types. The resultant enriched ontology was then mapped to the free text of peer-reviewed biological literature using the MMTx program producing a database of semantically enriched sentences.

A co-occurrence relation extraction algorithm was written to identify relationships between milk proteins and peptides, and other biological concepts, such as diseases or biological processes. Using these literature relation sets new hypotheses can be generated using the basic principle that if "*A* is linked to *B*", and if "*B* is linked to *C*" then we can infer an association between *A* and *C*. Filtering and downstream processing of the many generated relationships promotes significant interactions. These literature relations and hypotheses are integrated with biological data into the *MilkMine* database.

The *MilkMine* database is built upon on a generic data warehousing system, InterMine. This tool enabled the integration of traditional data types, such as protein sequence or structural data, from a variety of sources (*e.g.* UniProt). However, the standard InterMine model was also extended by the author to include other data sources (*e.g.* the Protein Data Bank) and to incorporate the output of the text-mining algorithm. This integration of otherwise disparate information allows more complex querying of the data, across many data types. For example, protein sequences are mapped to instances of the names, synonyms or symbols of the protein in text, therefore a raw fragment of amino acid sequence (*e.g.* a particular binding region) can be used to search the *MilkMine* database for literature information as well as the interactions and hypotheses of those

proteins that contain the sequence. The *MilkMine* resource is accessible online (www.bioinformatics.ed.ac.uk/milkmine) through a professional level query interface offering many features such as an interactive query builder, standard ready-to-run queries, bulk downloads and the ability to store user preferences and query histories.

Chapter 8 - Future work

A number of potential avenues to continue and extend the work covered in this thesis were identified and are detailed in this chapter.

8.1 Improvement of domain-related terminology extraction

One key improvement on the *MilkMine* system would be to improve the domain-related terminology generation. While the domain related terminology extraction (see Chapter 3) was of benefit, it did not generate an exhaustive list of important terms for milk researchers. As described in the Chapter 7, this could be improved by implementing the literature gradient technique over a larger number of citations. Also, the application of the literature gradient technique to another distinct domain of biology could be used to indicate the relative efficiency and portability of the theory.

Another improvement which could be made would be to increase the accuracy of the Named Entity Recognition (NER) stage in the process, particularly of proteins and genes. Incorporation of a protein/gene dictionary tagger would improve the recall and precision over the use of MMTx NER alone.

8.2 Scale-up of system to include full text articles

While it was recognised that literature abstracts were an appropriate and convenient input for this thesis due to their wide availability and high keyword density, full text articles contain a higher volume of information. However, a number of additional processes would need to be applied to scale-up the *MilkMine* program to use full text citations as input, such as collapsing fact redundancy and therefore although this was outwith the original scope of this thesis, it could be undertaken as a piece of future work.

8.3 Broaden scope of the MilkMine system

As was identified during evaluation of the *MilkMine* system (see Chapter 6) there was lower coverage of the protein chemistry aspects of milk proteins which were of interest to a number of research groups. Therefore, having established the system to cover more nutritional and physiological aspects of milk protein science a valuable further work would be to broaden the coverage of *MilkMine* to include more protein chemistry information.

8.4 Implementation of system architecture to another biological sub-domain

The work described in this thesis has attracted interest from a number of research groups, with specific requests for collaboration, namely from the Bruce German lab (University of California Davis, USA) and Donald Dunbar lab (University of Edinburgh, UK). Since the project has been undertaken with the aim of being portable to other biological sub-domains, much of the software and underlying structure developed by the author is directly transferable to other research areas. Thus two possible collaborative projects would be:

- Cardiovascular science discovery algorithm - Collaboration with Dr Donald Dunbar's group[1] to create a text-mining and database system for cardiovascular research. The large amount of literature and experimental information available on this research field would make an ideal candidate for such a resource.
- Nutrition science discovery algorithm - Collaboration with the Prof. German lab[2] to extend the current system to include all aspects of nutrition and metabolism.

[1] Centre for Cardiovascular Science and Centre for Inflammation Research, Queen's Medical Research Institute, University of Edinburgh, UK.
[2] Food Science and Technology, University of California Davis, USA.

Appendices

Appendix A: Summary of the results of the milk science research assessment of need questionnaire.

Area	Feature	Dick Fitzgerald	Kees de Kruif	Roger Clegg	Judy Hopkins	TOTAL
Molecular structure data	Secondary structure (e.g. percentage a-helix).	3	2	5	1	2.75
	Restriction enzyme sites.	3	5	3	1	3.00
	Disulphide bridge data.	4	5	3	1	3.25
	Structural motif information.	4	5	5	1	3.75
	Protein Data Bank structure.	4	5	5	1	3.75
	Protein family information.		5	4	1	3.33
	Denaturation properties.	5	5	2	5	4.25
	Thermostable properties.	5	5	2	5	4.25
	Iso-electric point.	5	5	3	2	3.75
	Protein-Protein interactions.	5	5	5	2	4.25
	Protein-Carbohydrate interactions.	4	5	2	2	3.25
	Protein-Lipid interactions.	4	3	2	2	2.75
	Protein-Nucleic acid interactions.	3	2	1	2	2.00
	Binding constants of ligands.	4	2	3	2	2.75
	Enzyme activities				3	3.00
	Other structural information.			5		5.00
Genomic information	Gene sequence.	3		5	1	3.00
	Exon-intron information.	2		5	1	2.67
	Chromosome location.	2		5	2	3.00
	Promoter sequence.	3		4	2	3.00
	Gene Ontology (GO) codes.	2		3	1	2.00

	Regulation of expression of lactation genes.	2		4	5	3.67
	Evolutionary information.	2		4	3	3.00
	Other genomic information.					
Dairy industry information	Milk yield information.	4	5	1	5	3.75
	Milk composition (protein, fat, carbohydrates).	5	5	1	5	4.00
	Effect of animal diet on milk composition.	4	2	1	5	3.00
	Milk storage properties (of milk and milk components)	4	2	1	5	3.00
	Processing properties.	5	2	1	5	3.25
	Gel-forming properties.	5	3	2	5	3.75
	Lactation (ee.g. information on regulation over a season).	3	1	2	5	2.75
	Genotype variant information.	5	1	4	5	3.75
	Genetic indices for cow breeding.	2	1	1		1.33
	Other dairy industry information.	4	1			2.50
Health related information	Neonatal growth.	3	3	1	5	3.00
	Passive immunity from milk.	4	4	2	5	3.75
	Antimicrobial properties.	5	4	3	5	4.25
	Physiological effects of milk (ee.g. immunity regulation).	5	4	3	5	4.25
	Calcium signalling pathways in the body.	4	3	4	4	3.75
	Antibodies raised in milk.	4	2	4	4	3.50

	Digestive enzyme activity on milk proteins.	5	3	2	4	3.50
	Bioactive peptides (digested milk proteins which then have a biological activity).	5	4	4	5	4.50
	Milk related diseases/disorders.	5	4	3	5	4.25
	Milk allergies.	5	4	3	5	4.25
	Chemical contaminants found in milk.	3	3	4	5	3.75
	Other health-related information.	5	1		5	3.67

Do you use genomic databases?	Yes	Yes	Yes	No
If so, which are your preferred DB?	ExPasy	PDB	Swissprot, Genbank, RIKEN	n/a
Where do you normally submit you queries?	sometimes NCBI also	n/a	MSA GeneStreamII (CNRS), CLUSTALW (EBI)	n/a
Preferred lit search engine?	Web of Sci>PubMed>Biosis	SciFinder	PubMed	Pubmed
Other web resources	-	-	EXPASY, Entrez, KinG, GNF Gene Expression Atlas, Sanger Centre, Protein Kinase resource	-
What mammalian species are you interested in	man, cow	cow	cow, rat, human, mouse	human
What other proteins are interesting	Proteinases,	n/a	GV protein ser/thr	bacteriostatic/immune

	peptidases	kinase, cAMP-dependent protein kinase, osteopontin	modulating proteins
Would microarray data be useful	Yes	yes	not immediately, perhaps in future
Any other features	- -	-	-
Text-mining uses	Milk protein structure, functionality and milk and health -	Heart/arterial disease, cell biology of secretion (constitutive secretory pathway)	Impact of meternal health on milk comosition and yield

Appendix B: Comparison of end-user text-mining systems (freely available, end-user systems only).

Feature	Chilibot (Chen and Sharp, 2004)	i-HOP (Hoffmann and Valencia, 2004)	EBIMed (Rebholz-Schuhmann et al., 2006)	Arrowsmith (Smalheiser and Swanson, 1998)	MilkMine
General					
Description	NLP-based relationship extraction program	Slick, fast, hyperlinked text	Relationship extractor, co-occurrence with LinkOuts	Co-occurrence, A-B-C link extraction	Domain specific, integrated, co-occurrence
Citation area coverage	Titles, abstract, MeSH terms	Titles, abstract	Titles, abstract (PDF full-text planned)	Titles, abstracts	title, abstract, MeSH terms
Relationship parameter	Phrase	Sentence	Sentence	Sentence	Sentence
Characterisation of relationship					
Use of MeSH terms in relation extraction	Partially to weight network nodes	No	No	Yes (semantic type information)	Yes
UMLS used	No	No	GO, drugs, species	No	Yes
MMTx used	No	No	No	No	Yes
NLP vs co-occurrence based	Both (shallow parse with CASS)	Co-occurrence (although records associative verbs)	Co-occurrence	Co-occurrence	Co-occurrence
NER	Pattern match collated dictionary (compiled from 6 genomic/proteomic databases) 113,503 unique symbols	UniProt, LocusLink	UniProt, GO, drugs, species. Acronym disambiguation	Term extraction with stop words removed	MMTx
Relation extraction evaluation	DIP known interactions	Manual checked ~3800 sentences	Manually checked ~300 sentences	Recreation of Swanson discoveries	Recreation of Swanson discoveries
Hypothesis generation	Yes	No	No	Yes	Yes

Feature	Chilibot (Chen and Sharp, 2004)	i-HOP (Hoffmann and Valencia, 2004)	EBIMed (Rebholz-Schuhmann et al., 2006)	Arrowsmith (Smalheiser and Swanson, 1998)	MilkMine
Domain specific	No	No	No	No	Yes (milk)
Microarray compatible	Yes, can superimpose expression levels	No	No	No	Potentially
Article limit	50 per term pair	Locally stored	10,000 for initial search	10,000 for initial search	1000 – 10000 citations
Speed	Minutes	Seconds	Seconds	Minutes	Rapid
Local storage vs on-the-fly	On-the-fly	Local storage	Local storage	On-the-fly	Local storage
# sentences	-	~12 million	-	-	~7.8 million
Literature sources used	PubMed, Google	PubMed	PubMed	PubMed	PubMed, CAB, OVID
End user interface					
Graphical representation of network	Yes, using AiSee	Yes	No	No	Yes (via download in cytoscape format – cytoscape plug-in being developed by InterMine project)
Graphical representation of links	Yes, using AiSee	Yes	Yes	Yes	Yes (via cytoscape)
Store user model	No	Yes	No	Yes	Yes
External link-outs	PubMed	PubMed, PubChem	PubMed, NCBI Taxonomy, NCBI drugs, UniProt, GO	PubMed	PubMed, UniProt, GO, PDB
API available	Perl script for remote searching				
Publication					
Lead author	Chen	Hoffmann	Rebholz-Schuhmann	Swanson	Edwards

Feature	Chilibot (Chen and Sharp, 2004)	i-HOP (Hoffmann and Valencia, 2004)	EBIMed (Rebholz-Schuhmann et al., 2006)	Arrowsmith (Smalheiser and Swanson, 1998)	MilkMine
Original publication year	2004	2005	2007	1996	
Publication (PMID)	15473905	16204114	17237098	8797484. The arrosmith project: 2005 status report, Dis. Sci. Proc	
Web interface	www.chilibot.net	www.ihop-net.org/UniPub/iHOP	www.ebi.ac.uk/Rebholz-srv/ebimed	http://arrowsmith.psych.uic.edu	www.bioinformatics.ed.ac.uk/milkmine

Appendix C: Semantic types used to categorise UMLS concepts within the UMLS metathesaurus.

Semantic Type Identifier	Semantic Type	Description
T001	Organism	Generally, a living individual, including all plants and animals.
T002	Plant	An organism having cellulose cell walls, growing by synthesis of inorganic substances, generally distinguished by the presence of chlorophyll, and lacking the power of locomotion. Plant parts are included here as well.
T003	Alga	A chiefly aquatic plant that contains chlorophyll, but does not form embryos during development and lacks vascular tissue.
T004	Fungus	A eukaryotic organism characterized by the absence of chlorophyll and the presence of a rigid cell wall. Included here are both slime molds and true fungi such as yeasts, molds, mildews, and mushrooms.
T005	Virus	An organism consisting of a core of a single nucleic acid enclosed in a protective coat of protein. A virus may replicate only inside a host living cell. A virus exhibits some but not all of the usual characteristics of living things.
T006	Rickettsia or Chlamydia	An organism intermediate in size and complexity between a virus and a bacterium, and which is parasitic within the cells of insects and ticks. Included here are all the chlamydias, also called "PLT" for psittacosis- lymphogranuloma venereum-trachoma.
T007	Bacterium	A small, typically one-celled, prokaryotic micro-organism.
T008	Animal	An organism with eukaryotic cells, and lacking stiff cell walls, plastids and photosynthetic pigments.
T009	Invertebrate	An animal which has no spinal column.
T010	Vertebrate	An animal which has a spinal column.
T011	Amphibian	A cold-blooded, smooth-skinned vertebrate which characteristically hatches as an aquatic larva, breathing by gills. When mature, the amphibian breathes with lungs.
T012	Bird	A vertebrate having a constant body temperature and characterized by the presence of feathers.
T013	Fish	A cold-blooded aquatic vertebrate characterized by fins and breathing by gills. Included here are fishes having either a bony skeleton, such as a perch, or a cartilaginous skeleton, such as a

		shark, or those lacking a jaw, such as a lamprey or hagfish.
T014	Reptile	A cold-blooded vertebrate having an external covering of scales or horny plates. Reptiles breathe by means of lungs and are generally egg-laying.
T015	Mammal	A vertebrate having a constant body temperature and characterized by the presence of hair, mammary glands and sweat glands.
T016	Human	Modern man, the only remaining species of the Homo genus.
T017	Anatomical Structure	A normal or pathological part of the anatomy or structural organization of an organism.
T018	Embryonic Structure	An anatomical structure that exists only before the organism is fully formed; in mammals, for example, a structure that exists only prior to the birth of the organism. This structure may be normal or abnormal.
T019	Congenital Abnormality	An abnormal structure, or one that is abnormal in size or location, present at birth or evolving over time as a result of a defect in embryogenesis.
T020	Acquired Abnormality	An abnormal structure, or one that is abnormal in size or location, found in or deriving from a previously normal structure. Acquired abnormalities are distinguished from diseases even though they may result in pathological functioning (ee.g., "hernias incarcerate").
T021	Fully Formed Anatomical Structure	An anatomical structure in a fully formed organism; in mammals, for example, a structure in the body after the birth of the organism.
T022	Body System	A complex of anatomical structures that performs a common function.
T023	Body Part, Organ, or Organ Component	A collection of cells and tissues which are localized to a specific area or combine and carry out one or more specialized functions of an organism. This ranges from gross structures to small components of complex organs. These structures are relatively localized in comparison to tissues.
T024	Tissue	An aggregation of similarly specialized cells and the associated intercellular substance. Tissues are relatively non-localized in comparison to body parts, organs or organ components.
T025	Cell	The fundamental structural and functional unit of living organisms.
T026	Cell Component	A part of a cell or the intercellular matrix, generally visible by light microscopy.
T028	Gene or Genome	A specific sequence, or in the case of the genome the complete sequence, of nucleotides along a molecule of DNA or RNA (in the case of some viruses) which represent the functional units of heredity.
T029	Body Location or Region	An area, subdivision, or region of the body demarcated for the purpose of topographical description.

T030	Body Space or Junction	An area enclosed or surrounded by body parts or organs or the place where two anatomical structures meet or connect.
T031	Body Substance	Extracellular material, or mixtures of cells and extracellular material, produced, excreted, or accreted by the body. Included here are substances such as saliva, dental enamel, sweat, and gastric acid.
T032	Organism Attribute	A property of the organism or its major parts.
T033	Finding	That which is discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient. The history of the presence of a disease is a 'Finding' and is distinguished from the disease itself.
T034	Laboratory or Test Result	The outcome of a specific test to measure an attribute or to determine the presence, absence, or degree of a condition.
T037	Injury or Poisoning	A traumatic wound, injury, or poisoning caused by an external agent or force.
T038	Biologic Function	A state, activity or process of the body or one of its systems or parts.
T039	Physiologic Function	A normal process, activity, or state of the body.
T040	Organism Function	A physiologic function of the organism as a whole, of multiple organ systems, or of multiple organs or tissues.
T041	Mental Process	A physiologic function involving the mind or cognitive processing.
T042	Organ or Tissue Function	A physiologic function of a particular organ, organ system, or tissue.
T043	Cell Function	A physiologic function inherent to cells or cell components.
T044	Molecular Function	A physiologic function occurring at the molecular level.
T045	Genetic Function	Functions of or related to the maintenance, translation or expression of the genetic material.
T046	Pathologic Function	A disordered process, activity, or state of the organism as a whole, of a body system or systems, or of multiple organs or tissues. Included here are normal responses to a negative stimulus as well as patholologic conditions or states that are less specific than a disease. Pathologic functions frequently have systemic effects.
T047	Disease or Syndrome	A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder.
T048	Mental or Behavioral Dysfunction	A clinically significant dysfunction whose major manifestation is behavioral or psychological. These dysfunctions may have identified or presumed biological etiologies or manifestations.

T049	Cell or Molecular Dysfunction	A pathologic function inherent to cells, parts of cells, or molecules.
T050	Experimental Model of Disease	A representation in a non-human organism of a human disease for the purpose of research into its mechanism or treatment.
T051	Event	A broad type for grouping activities, processes and states.
T052	Activity	An operation or series of operations that an organism or machine carries out or participates in.
T053	Behavior	Any of the psycho-social activities of humans or animals that can be observed directly by others or can be made systematically observable by the use of special strategies.
T054	Social Behavior	Behavior that is a direct result or function of the interaction of humans or animals with their fellows. This includes behavior that may be considered anti-social.
T055	Individual Behavior	Behavior exhibited by a human or an animal that is not a direct result of interaction with other members of the species, but which may have an effect on others.
T056	Daily or Recreational Activity	An activity carried out for recreation or exercise, or as part of daily life.
T057	Occupational Activity	An activity carried out as part of an occupation or job.
T058	Health Care Activity	An activity of or relating to the practice of medicine or involving the care of patients.
T059	Laboratory Procedure	A procedure, method, or technique used to determine the composition, quantity, or concentration of a specimen, and which is carried out in a clinical laboratory. Included here are procedures which measure the times and rates of reactions.
T060	Diagnostic Procedure	A procedure, method, or technique used to determine the nature or identity of a disease or disorder. This excludes procedures which are primarily carried out on specimens in a laboratory.
T061	Therapeutic or Preventive Procedure	A procedure, method, or technique designed to prevent a disease or a disorder, or to improve physical function, or used in the process of treating a disease or injury.
T062	Research Activity	An activity carried out as part of research or experimentation.
T063	Molecular Biology Research Technique	Any of the techniques used in the study of or the directed modification of the gene complement of a living organism.
T064	Governmental or Regulatory Activity	An activity carried out by officially constituted governments, or an activity related to the creation or enforcement of the rules or regulations governing some field of endeavor.
T065	Educational Activity	An activity related to the organization and provision of education.

T066	Machine Activity	An activity carried out primarily or exclusively by machines.
T067	Phenomenon or Process	A process or state which occurs naturally or as a result of an activity.
T068	Human-caused Phenomenon or Process	A phenomenon or process that is a result of the activities of human beings.
T069	Environmental Effect of Humans	A change in the natural environment that is a result of the activities of human beings.
T070	Natural Phenomenon or Process	A phenomenon or process that occurs irrespective of the activities of human beings.
T071	Entity	A broad type for grouping physical and conceptual entities.
T072	Physical Object	An object perceptible to the sense of vision or touch.
T073	Manufactured Object	A physical object made by human beings.
T074	Medical Device	A manufactured object used primarily in the diagnosis, treatment, or prevention of physiologic or anatomic disorders.
T075	Research Device	A manufactured object used primarily in carrying out scientific research or experimentation.
T077	Conceptual Entity	A broad type for grouping abstract entities or concepts.
T078	Idea or Concept	An abstract concept, such as a social, religious or philosophical concept.
T079	Temporal Concept	A concept which pertains to time or duration.
T080	Qualitative Concept	A concept which is an assessment of some quality, rather than a direct measurement.
T081	Quantitative Concept	A concept which involves the dimensions, quantity or capacity of something using some unit of measure, or which involves the quantitative comparison of entities.
T082	Spatial Concept	A location, region, or space, generally having definite boundaries.
T083	Geographic Area	A geographic location, generally having definite boundaries.
T085	Molecular Sequence	A broad type for grouping the collected sequences of amino acids, carbohydrates, and nucleotide sequences. Descriptions of these sequences are generally reported in the published literature and/or are deposited in and maintained by databanks such as GenBank, European Molecular Biology Laboratory (EMBL), National Biomedical Research Foundation (NBRF), or other sequence repositories.
T086	Nucleotide Sequence	The sequence of purines and pyrimidines in nucleic acids and polynucleotides. Included here are nucleotide-rich regions, conserved sequence, and DNA transforming region.

T087	Amino Acid Sequence	The sequence of amino acids as arrayed in chains, sheets, etc., within the protein molecule. It is of fundamental importance in determining protein structure.
T088	Carbohydrate Sequence	The sequence of carbohydrates within polysaccharides, glycoproteins, and glycolipids.
T089	Regulation or Law	An intellectual product resulting from legislative or regulatory activity.
T090	Occupation or Discipline	A vocation, academic discipline, or field of study, or a subpart of an occupation or discipline.
T091	Biomedical Occupation or Discipline	A vocation, academic discipline, or field of study related to biomedicine.
T092	Organization	The result of uniting for a common purpose or function. The continued existence of an organization is not dependent on any of its members, its location, or particular facility. Components or subparts of organizations are also included here. Although the names of organizations are sometimes used to refer to the buildings in which they reside, they are not inherently physical in nature.
T093	Health Care Related Organization	An established organization which carries out specific functions related to health care delivery or research in the life sciences.
T094	Professional Society	An organization uniting those who have a common vocation or who are involved with a common field of study.
T095	Self-help or Relief Organization	An organization whose purpose and function is to provide assistance to the needy or to offer support to those sharing similar problems.
T096	Group	A conceptual entity referring to the classification of individuals according to certain shared characteristics.
T097	Professional or Occupational Group	An individual or individuals classified according to their vocation.
T098	Population Group	An indivdual or individuals classified according to their sex, racial origin, religion, common place of living, financial or social status, or some other cultural or behavioral attribute.
T099	Family Group	An individual or individuals classified according to their family relationships or relative position in the family unit.
T100	Age Group	An individual or individuals classified according to their age.
T101	Patient or Disabled	An individual or individuals classified according to a disability, disease, condition or

	Group	treatment.
T102	Group Attribute	A conceptual entity which refers to the frequency or distribution of certain characteristics or phenomena in certain groups.
T103	Chemical	Compounds or substances of definite molecular composition. Chemicals are viewed from two distinct perspectives in the network, functionally and structurally. Almost every chemical concept is assigned at least two types, generally one from the structure hierarchy and at least one from the function hierarchy.
T104	Chemical Viewed Structurally	A chemical or chemicals viewed from the perspective of their structural characteristics. Included here are concepts which can mean either a salt, an ion, or a compound (ee.g., "Bromates" and "Bromides").
T109	Organic Chemical	The general class of carbon-containing compounds, usually based on carbon chains or rings, and also containing hydrogen (hydrocarbons), with or without nitrogen, oxygen, or other elements in which the bonding between elements is generally covalent.
T110	Steroid	One of a group of polycyclic, 17-carbon-atom, fused-ring compounds occurring both in natural and synthetic forms. Included here are naturally occurring and synthetic steroids, bufanolides, cardanolides, homosteroids, norsteroids, and secosteroids.
T111	Eicosanoid	An oxygenated metabolite from polyunsaturated 20 carbon fatty acids including lipoxygenase and cyclooxygenase products and their synthetic analogs. This includes the prostaglandins and thromboxanes.
T114	Nucleic Acid, Nucleoside, or Nucleotide	A complex compound of high molecular weight occurring in living cells. These are basically of two types, ribonucleic (RNA) and deoxyribonucleic (DNA) acids. Nucleic acids are made of nucleotides (nitrogen-containing base, a 5-carbon sugar, and one or more phosphate group) linked together by a phosphodiester bond between the 5' and 3' carbon atoms. Nucleosides are compounds composed of a purine or pyrimidine base (usually adenine, cytosine, guanine, thymine, uracil) linked to either a ribose or a deoxyribose sugar.
T115	Organophosphorus Compound	An organic compound containing phosphorus as a constituent. Included here are organic phosphinic, phosphonic and phosphoric acid derivatives and their thiophosphorus counterparts. Excluded are phospholipids, sugar phosphates, phosphoproteins, nucleotides, and nucleic acids.
T116	Amino Acid, Peptide, or Protein	Amino acids and chains of amino acids connected by peptide linkages.

T118	Carbohydrate	A generic term that includes monosaccharides, oligosaccharides, and polysaccharides as well as substances derived from monosaccharides by reduction of the carbonyl group (alditols), by oxidation of one or more terminal group to carboxylic acids, or by replacement of one or more hydroxy groups by a hydrogen atom, an amino group, a thiol group or similar heteroatomic groups. It also includes derivatives of these compounds. Included here are sugar phosphates. Excluded are glycolipids and glycoproteins.
T119	Lipid	An inclusive group of fat or fat-derived substances that are soluble in nonpolar solvents related to fatty acid esters, fatty alcohols, sterols, waxes, etc. Included in this group are the saponifiable lipids such as glycerides (fats and oils), essential (volatile) oils, and phospholipids.
T120	Chemical Viewed Functionally	A chemical viewed from the perspective of its functional characteristics or pharmacological activities.
T121	Pharmacologic Substance	A substance used in the treatment or prevention of pathologic disorders. This includes substances that occur naturally in the body and are administered therapeutically.
T122	Biomedical or Dental Material	A substance used in biomedicine or dentistry predominantly for its physical, as opposed to chemical, properties. Included here are biocompatible materials, tissue adhesives, bone cements, resins, toothpastes, etc.
T123	Biologically Active Substance	A generally endogenous substance produced or required by an organism, of primary interest because of its role in the biologic functioning of the organism that produces it.
T124	Neuroreactive Substance or Biogenic Amine	An endogenous substance whose activities affect or play an important role in the functioning of the nervous system. Included here are catecholamines, neuroregulators, neurophysins, etc.
T125	Hormone	In animals, a chemical usually secreted by an endocrine gland whose products are released into the circulating fluid. Hormones act as chemical messengers and regulate various physiologic processes such as growth, reproduction, metabolism, etc. They usually fall into two broad classes, steroid hormones and peptide hormones.
T126	Enzyme	A complex chemical, usually a protein, that is produced by living cells and which catalyzes specific biochemical reactions. There are six main types of enzymes: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases.
T127	Vitamin	A substance, usually an organic chemical complex, present in natural products or made synthetically, which is essential in the diet of man or other higher animals. Included here are

		vitamin precursors, provitamins, and vitamin supplements.
T129	Immunologic Factor	A biologically active substance whose activities affect or play a role in the functioning of the immune system.
T130	Indicator, Reagent, or Diagnostic Aid	A substance primarily of interest for its use in laboratory or diagnostic tests and procedures to detect, measure, examine, or analyze other chemicals, processes, or conditions.
T131	Hazardous or Poisonous Substance	A substance of concern because of its potentially hazardous or toxic effects. This would include most drugs of abuse, as well as agents that require special handling because of their toxicity.
T167	Substance	A material with definite or fairly definite chemical composition.
T168	Food	Any substance generally containing nutrients, such as carbohydrates, proteins, and fats, that can be ingested by a living organism and metabolized into energy and body tissue. Some foods are naturally occurring, others are either partially or entirely made by humans.
T169	Functional Concept	A concept which is of interest because it pertains to the carrying out of a process or activity.
T170	Intellectual Product	A conceptual entity resulting from human endeavor. Concepts assigned to this type generally refer to information created by humans for some purpose.
T171	Language	The system of communication used by a particular nation or people.
T184	Sign or Symptom	An observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation.
T185	Classification	A term or system of terms denoting an arrangement by class or category.
T190	Anatomical Abnormality	An abnormal structure, or one that is abnormal in size or location.
T191	Neoplastic Process	A new and abnormal growth of tissue in which the growth is uncontrolled and progressive. The growths may be malignant or benign.
T192	Receptor	A specific structure or site on the cell surface or within its cytoplasm that recognizes and binds with other specific molecules. These include the proteins on the surface of an immunocompetent cell that binds with antigens, or proteins found on the surface molecules that bind with hormones or neurotransmitters and react with other molecules that respond in a specific way.
T194	Archaeon	A member of one of the three domains of life, formerly called Archaebacteria under the taxon Bacteria, but now considered separate and distinct. Archaea are characterized by: 1) the

		presence of characteristic tRNAs and ribosomal RNAs; 2) the absence of peptidoglycan cell walls; 3) the presence of ether-linked lipids built from branched-chain subunits; and 4) their occurrence in unusual habitats. While archaea resemble bacteria in morphology and genomic organization, they resemble eukarya in their method of genomic replication.
T195	Antibiotic	A pharmacologically active compound produced by growing microorganisms which kill or inhibit growth of other microorganisms.
T196	Element, Ion, or Isotope	One of the 109 presently known fundamental substances that comprise all matter at and above the atomic level. This includes elemental metals, rare gases, and most abundant naturally occurring radioactive elements, as well as the ionic counterparts of elements (NA+, Cl-), and the less abundant isotopic forms. This does not include organic ions such as iodoacetate to which the type 'Organic Chemical' is assigned.
T197	Inorganic Chemical	Chemical elements and their compounds, excluding the hydrocarbons and their derivatives (except carbides, carbonates, cyanides, cyanates and carbon disulfide). Generally inorganic compounds contain ionic bonds. Included here are inorganic acids and salts, alloys, alkalies, and minerals.
T200	Clinical Drug	A pharmaceutical preparation as produced by the manufacturer. The name usually includes the substance, its strength, and the form, but may include the substance and only one of the other two items.
T201	Clinical Attribute	An observable or measurable property or state of an organism of clinical interest.
T203	Drug Delivery Device	A medical device that contains a clinical drug or drugs.

Appendix D: Medline citation format detailing the title, abstract, authors and MeSH terms.

PMID- 12888554
OWN - NLM
STAT- Medline
DA - 20031020
DCOM- 20040105
LR - 20061115
PUBM- Print-Electronic
IS - 0021-9258 (Print)
VI - 278
IP - 43
DP - 2003 Oct 24
TI - HAMLET interacts with histones and chromatin in tumor cell nuclei.
PG - 42131-5
AB - HAMLET is a folding variant of human alpha-lactalbumin in an active complex with
oleic acid. HAMLET selectively enters tumor cells, accumulates in their nuclei and induces apoptosis-like cell death. This study examined the interactions of HAMLET with nuclear constituents and identified histones as targets. HAMLET was
found to bind histone H3 strongly and to lesser extent histones H4 and H2B. The specificity of these interactions was confirmed using BIAcore technology and chromatin assembly assays. In vivo in tumor cells, HAMLET co-localized with histones and perturbed the chromatin structure; HAMLET was found associated with chromatin in an insoluble nuclear fraction resistant to salt extraction. In vitro, HAMLET bound strongly to histones and impaired their deposition on DNA. We
conclude that HAMLET interacts with histones and chromatin in tumor cell nuclei and propose that this interaction locks the cells into the death pathway by irreversibly disrupting chromatin organization.
AD - Institute of Laboratory Medicine, Section for Microbiology, Immunology and Glycobiology, Lund University, Solvegatan 23, 223 62 Lund, Sweden.
FAU - Duringer, Caroline
AU - Duringer C
FAU - Hamiche, Ali
AU - Hamiche A
FAU - Gustafsson, Lotta
AU - Gustafsson L
FAU - Kimura, Hiroshi
AU - Kimura H
FAU - Svanborg, Catharina
AU - Svanborg C

LA - eng
PT - Journal Article
PT - Research Support, Non-U.S. Gov't
DEP - 20030729
PL - United States
TA - J Biol Chem
JT - The Journal of biological chemistry
JID - 2985121R
RN - 0 (Antineoplastic Agents)
RN - 0 (Chromatin)
RN - 0 (Histones)
RN - 112-80-1 (Oleic Acid)
RN - 9013-90-5 (Lactalbumin)
SB - IM
MH - Active Transport, Cell Nucleus
MH - Antineoplastic Agents/metabolism/pharmacokinetics
MH - Cell Line, Tumor
MH - Cell Nucleus/metabolism/pathology
MH - Chromatin/*metabolism
MH - Histones/*metabolism
MH - Humans
MH - Lactalbumin/*metabolism/pharmacokinetics
MH - Oleic Acid
MH - Precipitation
MH - Protein Folding
MH - Protein Structure, Tertiary
EDAT- 2003/07/31 05:00
MHDA- 2004/01/06 05:00
PHST- 2003/07/29 [aheadofprint]
AID - 10.1074/jbc.M306462200 [doi]
AID - M306462200 [pii]
PST - ppublish
SO - J Biol Chem. 2003 Oct 24;278(43):42131-5. Epub 2003 Jul 29.

Appendix E: Milk related values (MRV) for journals which have at least 1,000 publications in Medline and with at least 100 assigned with the MeSH term, Milk[MH].

Milk related category	Journal	Frequency of Milk[MH] publications	Total publications	Milk related value (MRV)
Milk related (3) (ave. = 48.86%)	J Dairy Res	751	1298	57.85
	J Dairy Sci	3869	9702	39.87
Slightly related (2) (ave. = 6.73%)	J AOAC Int	194	2081	9.32
	Z Lebensm Unters Forsch	113	1301	8.68
	Int J Food Microbiol	257	3111	8.26
	J Pediatr Gastroentero l Nutr	438	5363	8.16
	J Food Prot	215	2701	7.96
	J Assoc Off Anal Chem	265	3453	7.67
	Nahrung	146	2012	7.25
	Br Vet J	164	2520	6.50
	Acta Vet Scand	145	2266	6.39
	Food Addit Contam	118	1952	6.04
	Vet Med (Praha)	129	2175	5.93
	Arch Latinoam Nutr	102	1720	5.93
	Vopr Pitan	312	5524	5.64
	Acta Paediatr Scand	220	4181	5.26
	Am J Clin Nutr	767	14746	5.20
	Br J Nutr	329	6417	5.12
	Vet Med Nauki	109	2130	5.11
Vaguely related (1)	J Appl Bacteriol	128	2689	4.76

(ave. = 2.10%)	Eur J Clin Nutr	156	3288	4.74
	J Appl Microbiol	111	2408	4.60
	Proc Nutr Soc	162	3712	4.36
	Rocz Panstw Zakl Hig	113	2705	4.17
	Zentralbl Veterinarmed A	112	2905	3.85
	J Anim Sci	418	12451	3.35
	Vet Microbiol	122	3670	3.32
	Lipids	179	5604	3.19
	Acta Paediatr	183	5767	3.17
	J Nutr	487	15642	3.11
	Biol Neonate	108	3468	3.11
	Pediatr Res	253	8338	3.03
	Dtsch Tierarztl Wochenschr	157	5320	2.95
	J Agric Food Chem	322	11162	2.88
	Tijdschr Diergeneeskd	142	4934	2.87
	Analyst	120	4252	2.82
	Nutr Rev	180	6799	2.64
	Health Phys	237	9550	2.48
	Am J Vet Res	382	15958	2.39
	Veterinariia	184	8061	2.28
	Monatsschr Kinderheilkd	147	6492	2.26
	Vet Rec	474	21942	2.16
	Res Vet Sci	115	5420	2.12
	Sci Total Environ	114	5457	2.08
	J Am Diet Assoc	176	8550	2.05
	Ann Allergy	111	5406	2.05
	Minerva Pediatr	218	10852	2.00

	Appl Microbiol	111	5554	1.99
	Chemosphere	102	5132	1.98
	Arch Dis Child	255	14169	1.79
	Bull Environ Contam Toxicol	163	9106	1.79
	J Pediatr	392	22589	1.73
	Pediatrics	402	25447	1.57
	J Am Vet Med Assoc	307	21569	1.42
	Adv Exp Med Biol	332	24872	1.33
	Am J Dis Child	110	9700	1.13
	Appl Environ Microbiol	213	19716	1.08
	J Chromatogr A	105	9878	1.06
	J Endocrinol	136	12998	1.04
	J Chromatogr	154	17782	0.86
	Gig Sanit	154	18066	0.85
	J Clin Microbiol	104	20260	0.51
	Arch Biochem Biophys	125	25027	0.49
	Biochim Biophys Acta	371	80017	0.46
	Biochem J	180	41430	0.43
	S Afr Med J	102	23634	0.43
	Lancet	478	113228	0.42
	Br Med J	179	44348	0.40
	Nature	259	79189	0.32
	J Biol Chem	261	133539	0.19
	Biochemistry	103	53226	0.19
	JAMA	108	57076	0.18
TOTAL	72 journals			

Appendix F: Unigram analysis results for milk[MH], milk protein[MH] and lactation[MH].

Key milk related MeSH term	Milk related category	Number of journals in corpus	Number of unigrams	Number of unique unigrams	% unique unigrams
Milk	Very related (5)	1163	1748873	34199	1.96
	Highly related (4)	1160	1868689	36926	1.98
	Related (3)	2	1798110	24577	1.37
	Slightly related (2)	11	1885566	37880	2.01
	Vaguely related (1)	46	1831209	61318	3.35
	General literature (0)	639	1682919	50459	3.00
	% variation		10.3	89.8	
Milk Protein	Very related (5)	1236	1660945	36797	2.22
	Highly related (4)	1212	1817733	41954	2.31
	Related (3)	2	1812000	24697	1.36
	Slightly related (2)	3	1863215	33939	1.82
	Vaguely related (1)	24	1744152	60540	3.47
	General literature (0)	639	1682919	50459	3.00
	% variation		11.5	86.6	
Lactation	Very related (5)	1291	1813842	28452	1.57
	Highly related (4)	1362	1932278	31839	1.65
	Related (3)	2	1812000	24697	1.36
	Slightly related (2)	7	2020645	33801	1.67
	Vaguely related (1)	19	1822509	48401	2.66
	General literature (0)	639	1682919	50459	3.00
	% variation		18.3	71.0	
Average	Very related (5)	1230	1741220	33149.3	1.90
	Highly related (4)	1244.7	1872900	36906.3	1.97
	Related (3)	2	1807370	24657.0	1.36
	Slightly related (2)	7	1923142	35206.7	1.83
	Vaguely related (1)	29.7	1799290	56753.0	3.15
	General literature (0)	639	1682919	50459.0	3.00
	% variation		13.3	81.2	

Appendix G Source vocabularies used to create the customised 2006AB UMLS dataset for the MilkMine database.

Source vocabulary	Dataset
AI/RHEUM,	1993
Alcohol and Other Drug Thesaurus,	2000
CRISP Thesaurus,	2006
Common Terminology Criteria for Adverse Events,	2003
Gene Ontology,	2006_01_20
HUGO Gene Nomenclature,	2005_04
ICD-9-CM,	2006
Medline	(1996-2000)
McMaster University Epidemiology Terms,	1992
Medical Dictionary for Regulatory Activities Terminology	(MedDRA)
Medline	(2001-2006)
MedlinePlus Health Topics	2004_08_14
Medical Subject Headings	2006_2006_02_06
UMLS Metathesaurus	2006AB
Metathesaurus CPT Hierarchical Terms,	2006
Metathesaurus FDA National Drug Code Directory,	2005_06_30
Metathesaurus additional entry terms for ICD-9-CM,	2006
Metathesaurus Version of Minimal Standard Terminology Digestive	2001
NCBI Taxonomy,	2006_01_04
NCI modified Common Terminology Criteria for Adverse Events v3.0	NCI-CTCAEV3
NCI Thesaurus,	2004_11_17
National Library of Medicine Medline Data	NLM-MED
RXNORM Project, META2006AA Full Update	2006_03_14,
SNOMED-2, 2	SNM2
SNOMED International,	1998
SNOMED Clinical Terms	2006_01_31
University of Washington Digital Anatomist,	1.7.3

Appendix H: Semantic types used as Primary concepts for literature relation analysis within the *MilkMine* text-mining algorithm.

UMLS type ID	Discovery Semantic Type
T023	Body Part, Organ, or Organ Component
T025	Cell
T028	Gene or Genome
T031	Body Substance
T032	Organism Attribute
T033	Finding
T034	Laboratory or Test Result
T038	Biologic Function
T039	Physiologic Function
T041	Mental Process
T042	Organ or Tissue Function
T043	Cell Function
T044	Molecular Function
T046	Pathologic Function
T047	Disease or Syndrome
T048	Mental or Behavioral Dysfunction
T059	Laboratory Procedure
T067	Phenomenon or Process
T068	Human-caused Phenomenon or Process
T079	Temporal Concept
T080	Qualitative Concept
T098	Population Group
T109	Organic Chemical
T110	Steroid
T111	Eicosanoid
T116	Amino Acid, Peptide, or Protein
T118	Carbohydrate
T119	Lipid
T120	Chemical Viewed Functionally
T121	Pharmacologic Substance
T123	Biologically Active Substance
T124	Neuroreactive Substance or Biogenic Amine
T125	Hormone
T126	Enzyme
T127	Vitamin
T129	Immunologic Factor
T130	Indicator, Reagent, or Diagnostic Aid
T169	Functional Concept
T184	Sign or Symptom
T185	Classification
T192	Receptor
T196	Element, Ion, or Isotope
T200	Clinical Drug

Appendix I: MilkMine interface tutorial

The *MilkMine* interface tutorial is a large document and therefore has not been included here as an Appendix. The tutorial is however available online at http://www.bioinformatics.ed.ac.uk/milkmine/tutorial.doc.

Publications

Poster Presentations

BioSysBio 2005, Edinburgh (poster abstract)

Text-mining, milk proteins and nutraceutical potential – the MilkER project

S*e.g.* Edwards [1]*, B. Webber [1], C. Holt [2], L. Sawyer [1]
[1]*University of Edinburgh, UK;* [2]*Hannah Research Institute, UK*
s0460205@sms.ed.ac.uk

The vast amount of literature on milk proteins and genes, and bioactive milk-protein derived peptides cries out for a single informatics resource to focus development of research in the food, health and medical industries. The *milkER* (Milk Extraction Resource) project aims to provide this. The database contains milk protein and gene sequence information, ligand binding data, bioactive peptide data and protein-protein/disease interaction data for many mammalian species. In addition to a milk literature interface, we aim to include data on the effects of milk composition on growth and health, enzymatic properties of milk proteins, and also proteomic and microarray data.

As well as data collation, *milkER* also aims to perform text-mining on milk literature allowing discovery of novel functional relationships among milk proteins under physiological and processing conditions, leading to potential health and manufacturing benefits. This will be focused on the interactions of milk proteins physiologically with respect to positive *and* negative effects in mother, child and consumer. In comparison with labour-based research, conceptual research is more cost-effective and *milkER* will provide high throughput analysis of the milk literature. The *milkER* website is online at **www.milkER.org.uk**.

Keywords: Milk, Informatics, Database, Text-mining

IMGC Milk Genomics 2007, Brussels (poster and abstract)

MilkMine - milking the dairy literature

S*e.g.* Edwards [1], B. Webber [1], C. Holt [2], L. Sawyer [1]
[1]*University of Edinburgh, UK;* [2]*Hannah Research Institute, UK*
stephen@milker.org.uk

The *MilkMine* project aims to bring together data from empirical research with emerging text-mining techniques. The vast amount of literature on milk proteins and their genes, as well as bioactive milk-protein derived peptides means that researchers often struggle to keep up with the information expansion and may not recognise vital biological links, particularly those that extend outwith their area of expertise. *MilkMine* is an attempt to provide a single informatics resource to help researchers mine this information mountain and to help focus development of research in the food, health and medical industries.

MilkMine is based on a generic data warehousing system, InterMine, enabling the integration of many traditional data types, such as UniProt protein data, genomes and comparative genomic data, as well as protein interaction data. During the data

integration step, links are made between the data types, for example, finding orthologues for a given milk protein. Already a useful tool, this structure has been extended here to make use of text-mining techniques.

Milk and lactation specific terminology was identified to complement the underlying UMLS metathesaurus. This valuable, milk-related ontology can be used to identify biological concepts in free text and has been applied to milk literature contained in the PubMed database. The resultant semantically enriched literature allows the derivation of relationships between milk proteins, genes and peptides, as well as other biological concepts, such as diseases or biological processes. In this way we can create new hypotheses using the basic principle that if "*A* is linked to *B*", and if "*B* is linked to *C*" then we can infer an association between *A* and *C*. Filtering and downstream processing of the many generated relationships help to reduce the number of non-significant interactions and improve the scoring of novel ones, for example, by removal of known associations where an *A* to *C* link is widely known.

The *MilkMine* resource is accessible online through an excellent query interface allowing users to generate and perform complex queries across the data model. In comparison with labour-based research, conceptual research is more cost-effective and it is hoped that this system will lead to the discovery of novel functional relationships among milk proteins under physiological and processing conditions, leading to manufacturing and health benefits for mother, child and consumer. These hypotheses can then be verified in the laboratory.

Keywords: Milk, Informatics, Database, Text-mining, knowledge discovery.

(417 Words)

MilkMine - an integrated database of milk knowledge

S.G.Edwards[1], B.Webber[2], C.Holt and L.Sawyer[1].

[1]ISMB, The University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh, UK.

[2] School of Informatics, The University of Edinburgh, King's Buildings, West Mains Road, Edinburgh, UK.

Introduction

The ever-increasing amount of data and literature on milk proteins, genes and bioactive peptides means that it can be a struggle for researchers to keep on top of it and relate it to other fields of research. *MilkMine* is an attempt to provide a single resource to help researchers make maximum use of this information mountain in an easy-to-use web database.

Data integration

MilkMine integrates many traditional data types with information from the literature:

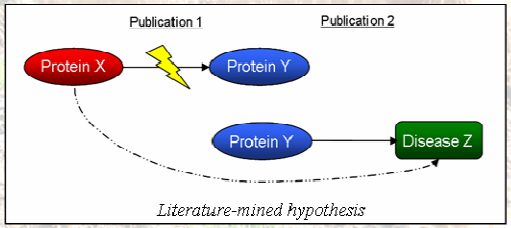
- **Proteins** – extensive collection of mammalian proteins, including those present in milk
- **Protein Structures** – milk protein crystal structures (protein structure viewer available on website)
- **Protein-Protein Interactions** – collection of high throughput experiment data
- **Genes** - extensive collection of mammalian genes, including those whose products are present in milk
- **Annotation** – functional annotation for proteins and genes in the database
- **Bioactive milk peptides** – collection of peptides from milk proteins with identified bioactivity
- **Milk allergens** – collection of allergenic milk components
- **MilkMine library** – extensive collection of literature on milk, colostrum, lactation and milk proteins and peptides
- **Literature-mining** – summarises and analyses milk literature (see next panel)

Literature-mining

Literature-mining uses computer programs to analyse many scientific publications at once and spot significant relationships between biological concepts, for example between proteins, peptides or diseases:

Publication 1 states that:
"***Protein X*** **has been shown to activate** ***Protein Y*****.**"

Publication 2 states that:
"***Protein Y*** **is part of the pathway of** ***Disease Z*****.**"



Literature-mined hypothesis

Therefore, we can make a hypothesis that there may be a relationship between ***Protein X*** and ***Disease Z*** (see Figure). MilkMine performs this analysis on milk literature and identifies any **little known** or **unstudied hypotheses** which represent those most likely to be interesting.

These hypotheses are stored in the MilkMine database with links to their context in the literature and can be searched through the MilkMine website at:

www.bioinformatics.ed.ac.uk/MilkMine

Database features

- The database can be searched by:
  - Quick search using name, synonym or identifiers.
  - Templates, a collection of commonly used database searches.
  - Create your own search!
- Perform searches using (uploadable) lists
- Save queries and results
- Download data in a variety of formats
- Website tour and tutorial

Possible queries

Complex queries can be performed on the data in the MilkMine database, such as:

- Is there a connection between beta-casomorphin and Sudden Infant Death Syndrome?
- What is caseinomacropeptide?
- Show me all sentences in the scientific literature which refer to any milk protein that has the specific allergenic epitope 'KKILDKVGIN'.
- Search milk literature by organism or population group (e.g. human premature babies)

Acknowledgements

The project is funded by a BBSRC studentship. Thanks go to Dr Alistair Kerr, the InterMine team (Cambridge) and Dr Sophia Ananiadou for technical support and advice.

Beta-lactoglobulin book chapter

Edwards S., Sawyer L: **Functional aspects of β-lactoglobulin, Major Urinary Protein and Odorant-Binding Protein**. *Lipocalins*, Landes Biosciences 2005.

References

UniProt. (2008) The universal protein resource (UniProt). *Nucleic Acids Res,* 36**,** D190-5.

ADJAYE, J., HERWIG, R., HERRMANN, D., WRUCK, W., BENKAHLA, A., BRINK, T. C., NOWAK, M., CARNWATH, J. W., HULTSCHIG, C., NIEMANN, H. & LEHRACH, H. (2004) Cross-species hybridisation of human and bovine orthologous genes on high density cDNA microarrays. *BMC Genomics,* 5**,** 83.

APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BIRNEY, E., BISWAS, M., BUCHER, P., CERUTTI, L., CORPET, F., CRONING, M. D., DURBIN, R., FALQUET, L., FLEISCHMANN, W., GOUZY, J., HERMJAKOB, H., HULO, N., JONASSEN, I., KAHN, D., KANAPIN, A., KARAVIDOPOULOU, Y., LOPEZ, R., MARX, B., MULDER, N. J., OINN, T. M., PAGNI, M., SERVANT, F., SIGRIST, C. J. & ZDOBNOV, E. M. (2000) InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics,* 16**,** 1145-50.

ARONSON, A. R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp***,** 17-21.

ASHBURNER, M. & DRYSDALE, R. (1994) FlyBase--the Drosophila genetic database. *Development,* 120**,** 2077-9.

AZUMA, N. & YAMAUCHI, K. (1991) Identification of alpha-lactalbumin and beta-lactoglobulin in cynomolgus monkey (Macaca fascicularis) milk. *Comp Biochem Physiol B,* 99**,** 917-21.

BAND, M. R., LARSON, J. H., REBEIZ, M., GREEN, C. A., HEYEN, D. W., DONOVAN, J., WINDISH, R., STEINING, C., MAHYUDDIN, P., WOMACK, J. E. & LEWIN, H. A. (2000) An ordered comparative map of the cattle and human genomes. *Genome Res,* 10**,** 1359-68.

BARILLET, F., ARRANZ, J. J. & CARTA, A. (2005) Mapping quantitative trait loci for milk production and genetic polymorphisms of milk proteins in dairy sheep. *Genet Sel Evol,* 37**,** S109-S123.

BAUMGARTNER, W. A., JR., COHEN, K. B. & HUNTER, L. (2008) An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. *J Biomed Discov Collab,* 3**,** 1.

BELIVEAU, R. & GINGRAS, D. (2007) Role of nutrition in preventing cancer. *Can Fam Physician,* 53**,** 1905-11.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28**,** 235-42.

BESSELINK, M. G., TIMMERMAN, H. M., BUSKENS, E., NIEUWENHUIJS, V. B., AKKERMANS, L. M. & GOOSZEN, H. G. (2004) Probiotic prophylaxis in patients with predicted severe acute pancreatitis (PROPATRIA): design and rationale of a double-blind, placebo-controlled randomised multicenter trial [ISRCTN38327949]. *BMC Surg,* 4**,** 12.

BLASCHKE, C., HIRSCHMAN, L. & VALENCIA, A. (2002) Information extraction in molecular biology. *Brief Bioinform,* 3**,** 154-65.

BOBE, G., BEITZ, D. C., FREEMAN, A. E. & LINDBERG, G. L. (1999) Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. *J Dairy Sci,* 82**,** 2797-804.

BOEHM, G. & STAHL, B. (2007) Oligosaccharides from milk. *J Nutr,* 137**,** 847S-9S.

BREW, K. (1969) Secretion of alpha-lactalbumin into milk and its relevance to the organization and control of lactose synthetase. *Nature,* 222**,** 671-2.

BUHIMSCHI, C. S. (2004) Endocrinology of lactation. *Obstet Gynecol Clin North Am,* 31**,** 963-79, xii.

BUNESCU, R., GE, R., KATE, R. J., MARCOTTE, E. M., MOONEY, R. J., RAMANI, A. K. & WONG, Y. W. (2005) Comparative experiments on learning information extractors for proteins and their interactions. *Artif Intell Med,* 33**,** 139-55.

CHANG, J. T., SCHUTZE, H. & ALTMAN, R. B. (2004) GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics,* 20**,** 216-25.

CHARLWOOD, J., HANRAHAN, S., TYLDESLEY, R., LANGRIDGE, J., DWEK, M. & CAMILLERI, P. (2002) Use of proteomic methodology for the characterization of human milk fat globular membrane proteins. *Anal Biochem,* 301**,** 314-24.

CHEN, H. & SHARP, B. M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics,* 5**,** 147.

CHUN, H. W., TSURUOKA, Y., KIM, J. D., SHIBA, R., NAGATA, N., HISHIKI, T. & TSUJII, J. (2006) Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. *BMC Bioinformatics,* 7 Suppl 3**,** S4.

CLARE, D. A. & SWAISGOOD, H. E. (2000) Bioactive milk peptides: a prospectus. *J Dairy Sci,* 83**,** 1187-95.

CLARKSON, R. W. & WATSON, C. J. (2003) Microarray analysis of the involution switch. *J Mammary Gland Biol Neoplasia,* 8**,** 309-19.

COHEN, K. B. & HUNTER, L. (2004) *Natural language processing and systems biology.*

COHEN, K. B. & HUNTER, L. (2008) Getting started in text mining. *PLoS Comput Biol,* 4**,** e20.

COLE, R. J. & BRUZA, P. D. (2005) A bare bones approach to literature-based discovery: An analysis of the Raynaud's/fish-oil and migraine-magnesium discoveries in semantic space. *Discovery Science, Proceedings.*

CORNEY, D. P., BUXTON, B. F., LANGDON, W. B. & JONES, D. T. (2004) BioRAT: extracting biological information from full-length papers. *Bioinformatics,* 20**,** 3206-13.

DANBY, F. W. (2005) Acne and milk, the diet myth, and beyond. *J Am Acad Dermatol,* 52**,** 360-2.

DELLAIRE, G., FARRALL, R. & BICKMORE, W. A. (2003) The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucleic Acids Res,* 31**,** 328-30.

DEMMER, J., STASIUK, S. J., GRIGOR, M. R., SIMPSON, K. J. & NICHOLAS, K. R. (2001) Differential expression of the whey acidic protein gene during lactation in the brushtail possum (Trichosurus vulpecula). *Biochim Biophys Acta,* 1522**,** 187-94.

DIGIACOMO, R. A., KREMER, J. M. & SHAH, D. M. (1989) Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am J Med,* 86**,** 158-64.

DING, J., BERLEANT, D., NETTLETON, D. & WURTELE, E. (2002) Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput***,** 326-37.

DZIUBA, J., MINKIEWICZ, P., NALECZ, D. & IWANIAK, A. (1999) Database of biologically active peptide sequences. *Nahrung,* 43**,** 190-5.

EDWARDS, S., WEBBER, B., HOLT, C. & SAWYER, L. (2005) Text-mining, milk proteins and nutraceutical potential - the MilkER project (Poster). *BioSysBio.* Suppl 3 ed. Edinburgh, UK, BMC Bioinformatics.

ELLIOT, WASMUTH, BIBBY & HILL (1997) The role of beta-casein variants in the induction of insulin-dependent diabetes in the non-obese diabetic mouse and humans. *IDF - Milk Protein Polymorphism seminar.* New Zealand.

ELLIOTT, R. B., HARRIS, D. P., HILL, J. P., BIBBY, N. J. & WASMUTH, H. E. (1999) Type I (insulin-dependent) diabetes mellitus and cow milk: casein variant consumption. *Diabetologia,* 42**,** 292-6.

ELSEN, J. M. (2005) Foreword to the international workshop on major genes and QTL in sheep and goats. *Genet Sel Evol,* 37 Suppl 1**,** II.

FANARO, S., CHIERICI, R., GUERRINI, P. & VIGI, V. (2003) Intestinal microflora in early infancy: composition and development. *Acta Paediatr Suppl,* 91**,** 48-55.

FARKYE, N. Y. (2003) *Other enzymes*.

FARRELL, H. M., JR., JIMENEZ-FLORES, R., BLECK, G. T., BROWN, E. M., BUTLER, J. E., CREAMER, L. K., HICKS, C. L., HOLLAR, C. M., NG-KWAI-HANG, K. F. & SWAISGOOD, H. E. (2004) Nomenclature of the proteins of cows' milk--sixth revision. *J Dairy Sci,* 87**,** 1641-74.

FARRELL, H. M. & THOMPSON, M. P. (1990) Beta-Lactoglobulin And Alpha-Lactalbumin As Potential Modulators Of Mammary Cellular-Activity - A Ca2+-Responsive Model System Using Acid Phosphoprotein Phosphatases. *Protoplasma,* 159**,** 157-167.

FLORISA, R., RECIO, I., BERKHOUT, B. & VISSER, S. (2003) Antibacterial and antiviral effects of milk proteins and derivatives thereof. *Curr Pharm Des,* 9**,** 1257-75.

FLOWER, D. R., NORTH, A. C. T. & SANSOM, C. E. (2000) The lipocalin protein family: structural and sequence overview. *Biochimica Et Biophysica Acta-Protein Structure And Molecular Enzymology,* 1482**,** 9-24.

FOLCH, J. M., COLL, A., HAYES, H. C. & SANCHEZ, A. (1996) Characterization of a caprine beta-lactoglobulin pseudogene, identification and chromosomal localization by in situ hybridization in goat, sheep and cow. *Gene,* 177**,** 87-91.

FOX, P. F. (2003a) *Advanced Dairy Chemistry*, Kluwer Academic.

FOX, P. F. (2003b) *Indigenous enzymes in milk*.

FUGLSANG, A., NILSSON, D. & NYBORG, N. C. (2003) Characterization of new milk-derived inhibitors of angiotensin converting enzyme in vitro and in vivo. *J Enzyme Inhib Med Chem,* 18**,** 407-12.

GALPERIN, M. Y. (2008) The Molecular Biology Database Collection: 2008 update. 36**,** D2-4.

GANFORNINA, M. D., GUTIERREZ, G., BASTIANI, M. & SANCHEZ, D. (2000) A phylogenetic analysis of the lipocalin protein family. *Mol Biol Evol,* 17**,** 114-26.

GARTNER, L. M., MORTON, J., LAWRENCE, R. A., NAYLOR, A. J., O'HARE, D., SCHANLER, R. J. & EIDELMAN, A. I. (2005) Breastfeeding and the use of human milk. *Pediatrics,* 115**,** 496-506.

GERMAN, J. B., SCHANBACHER, F. L., LÖNNERDAL, B., MEDRANO, J. F., MCGUIRE, M. A., MCMANAMAN, J. L., ROCKE, D. M., SMITH, T. P., NEVILLE, M. C., DONNELLY, P., LANGE, M. & WARD, R. (2006) International milk genomics consortium. *Trends in Food Science & Technology,* 17**,** 656-661.

GORDON, M. D. & LINDSAY, R. K. (1996) Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal Of The American Society For Information Science,* 47**,** 116-128.

HAMBRAEUS, L. & LONNERDAL, B. (2003) *Nutritional aspects of milk proteins*.

HANSON, L. A. & KOROTKOVA, M. (2002) The role of breastfeeding in prevention of neonatal infection. *Semin Neonatol,* 7**,** 275-81.

HARTMANN, R. & MEISEL, H. (2007) Food-derived peptides with biological activity: from research to food applications. *Current Opinion In Biotechnology,* 18**,** 163-169.

HE, X. & ZHANG, J. (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet,* 2**,** e88.

HERMJAKOB, H., MONTECCHI-PALAZZI, L., LEWINGTON, C., MUDALI, S., KERRIEN, S., ORCHARD, S., VINGRON, M., ROECHERT, B., ROEPSTORFF, P., VALENCIA, A., MARGALIT, H., ARMSTRONG, J., BAIROCH, A., CESARENI, G., SHERMAN, D. & APWEILER, R. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res,* 32**,** D452-5.

HIRSCHMAN, L. & BLASCHKE, C. (2006) Evaluation of text mining in biology. IN ANANIADOU, S. & MCNAUGHT, J. (Eds.) *Text mining for Biology and Biomedicine.* 1 ed. London, Artech House.

HOFFMANN, R. & VALENCIA, A. (2004) A gene network for navigating the literature. *Nat Genet,* 36**,** 664.

HOFFMANN, R. & VALENCIA, A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics,* 21 Suppl 2**,** ii252-ii258.

HOPPE, C., MOLGAARD, C., JUUL, A. & MICHAELSEN, K. F. (2004a) High intakes of skimmed milk, but not meat, increase serum IGF-I and IGFBP-3 in eight-year-old boys. *Eur J Clin Nutr,* 58**,** 1211-6.

HOPPE, C., MOLGAARD, C., VAAG, A., BARKHOLT, V. & MICHAELSEN, K. F. (2005) High intakes of milk, but not meat, increase s-insulin and insulin resistance in 8-year-old boys. *Eur J Clin Nutr,* 59**,** 393-8.

HOPPE, C., UDAM, T. R., LAURITZEN, L., MOLGAARD, C., JUUL, A. & MICHAELSEN, K. F. (2004b) Animal protein intake, serum insulin-like growth factor I, and growth in healthy 2.5-y-old Danish children. *Am J Clin Nutr,* 80**,** 447-52.

HUANG, M., ZHU, X., HAO, Y., PAYAN, D. G., QU, K. & LI, M. (2004) Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics,* 20**,** 3604-12.

HUBBARD, T., BARKER, D., BIRNEY, E., CAMERON, G., CHEN, Y., CLARK, L., COX, T., CUFF, J., CURWEN, V., DOWN, T., DURBIN, R., EYRAS, E., GILBERT, J., HAMMOND, M., HUMINIECKI, L., KASPRZYK, A., LEHVASLAIHO, H., LIJNZAAD, P., MELSOPP, C., MONGIN, E., PETTETT, R., POCOCK, M., POTTER, S., RUST, A., SCHMIDT, E., SEARLE, S., SLATER, G., SMITH, J., SPOONER, W., STABENAU, A., STALKER, J., STUPKA, E., URETA-VIDAL, A., VASTRIK, I. & CLAMP, M. (2002) The Ensembl genome database project. *Nucleic Acids Res,* 30**,** 38-41.

HUNTER, L. & COHEN, K. B. (2006) Biomedical Language Processing: What's Beyond PubMed? *Mol Cell,* 21**,** 589-94.

JAYAT, D., GAUDIN, J. C., CHOBERT, J. M., BUROVA, T. V., HOLT, C., MCNAE, I., SAWYER, L. & HAERTLE, T. (2004) A recombinant C121S mutant of bovine beta-lactoglobulin is more susceptible to peptic digestion and to denaturation by reducing agents and heating. *Biochemistry,* 43**,** 6312-21.

JENSSEN, T. K., LAEGREID, A., KOMOROWSKI, J. & HOVIG, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet,* 28**,** 21-8.

JOSE, H., VADIVUKARASI, T. & DEVAKUMAR, J. (2007) Extraction of Protein Interaction Data: A Comparative Analysis of Methods in Use. *EURASIP Journal on Bioinformatics and Systems Biology,* 2007.

KAILA, M., ARVILOMMI, H., SOPPI, E., LAINE, S. & ISOLAURI, E. (1994) A prospective study of humoral immune responses to cow milk antigens in the first year of life. *Pediatr Allergy Immunol,* 5**,** 164-9.

KAMINSKI, S., AHMAN, A., RUSC, A., WOJCIK, E. & MALEWSKI, T. (2005) MilkProtChip--a microarray of SNPs in candidate genes associated with milk protein biosynthesis--development and validation. *J Appl Genet,* 46**,** 45-58.

KAYSER, H. & MEISEL, H. (1996) Stimulation of human peripheral blood lymphocytes by bioactive peptides derived from bovine milk proteins. *FEBS Lett,* 383**,** 18-20.

KIM, H. H. & JIMENEZ-FLORES, R. (1994) Comparison of milk proteins using preparative isoelectric focusing followed by polyacrylamide gel electrophoresis. *J Dairy Sci,* 77**,** 2177-90.

KIM, J. D., OHTA, T., TATEISI, Y. & TSUJII, J. (2003) GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics,* 19 Suppl 1**,** i180-2.

KOCH, G., WIEDEMANN, K., DREBES, E., ZIMMERMANN, W., LINK, G. & TESCHEMACHER, H. (1988) Human beta-casomorphin-8 immunoreactive material in the plasma of women during pregnancy and after delivery. *Regul Pept,* 20**,** 107-17.

KONTOPIDIS, G., HOLT, C. & SAWYER, L. (2002) The ligand-binding site of bovine beta-lactoglobulin: evidence for a function? *J Mol Biol,* 318**,** 1043-55.

KONTOPIDIS, G., HOLT, C. & SAWYER, L. (2004) Invited review: beta-lactoglobulin: binding properties, structure, and function. *J Dairy Sci,* 87**,** 785-96.

KOSTOFF, R. N., BLOCK, J. A., STUMP, J. A. & PFEIL, K. M. (2004) Information content in Medline record fields. *International Journal Of Medical Informatics,* 73**,** 515-527.

KRALLINGER, M., ERHARDT, R. A. & VALENCIA, A. (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discov Today,* 10**,** 439-45.

KRALLINGER, M. & VALENCIA, A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol,* 6**,** 224.

KRISSANSEN, G. W. (2007) Emerging health properties of whey proteins and their clinical implications. *J Am Coll Nutr,* 26**,** 713S-23S.

KUMOSINSKI, T. F., BROWN, E. M. & FARRELL, H. M., JR. (1993) Three-dimensional molecular modeling of bovine caseins: an energy-minimized beta-casein structure. *J Dairy Sci,* 76**,** 931-45.

LACAZETTE, E., GACHON, A. M. & PITIOT, G. (2000) A novel human odorant-binding protein gene family resulting from genomic duplicons at 9q34: differential expression in the oral and genital spheres. *Hum Mol Genet,* 9**,** 289-301.

LAWLOR, D. A., EBRAHIM, S., TIMPSON, N. & DAVEY SMITH, G. (2005) Avoiding milk is associated with a reduced risk of insulin resistance and the metabolic syndrome: findings from the British Women's Heart and Health Study. *Diabet Med,* 22**,** 808-11.

LEMAY, D. G., NEVILLE, M. C., RUDOLPH, M. C., POLLARD, K. S. & GERMAN, J. B. (2007) Gene regulatory networks in lactation: identification of global principles using bioinformatics. *BMC Syst Biol,* 1**,** 56.

LEMKIN, P. F., THORNWALL, G. C., WALTON, K. D. & HENNIGHAUSEN, L. (2000) The microarray explorer tool for data mining of cDNA microarrays: application for the mammary gland. *Nucleic Acids Res,* 28**,** 4452-9.

LESER, U. & HAKENBERG, J. (2005) What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform,* 6**,** 357-69.

LEWIN, H. A. (2003) The future of cattle genome research: The beef is here. *Cytogenic and Genome Research,* 102**,** 10 - 15.

LINDSAY, R. K. & GORDON, M. D. (1999) Literature-based discovery by lexical statistics. *Journal Of The American Society For Information Science,* 50**,** 574-587.

LIVNEY, Y. D., SCHWAN, A. L. & DALGLEISH, D. G. (2004) A study of beta-casein tertiary structure by intramolecular crosslinking and mass spectrometry. *J Dairy Sci,* 87**,** 3638-47.

LONNERDAL, B. (1985) Biochemistry and physiological function of human milk proteins. *Am J Clin Nutr,* 42**,** 1299-317.

LUM, L. S., DOVC, P. & MEDRANO, J. F. (1997) Polymorphisms of bovine beta-lactoglobulin promoter and differences in the binding affinity of activator protein-2 transcription factor. *J Dairy Sci,* 80**,** 1389-97.

MACFARLANE, A. J. & STOVER, P. J. (2007) Convergence of genetic, nutritional and inflammatory factors in gastrointestinal cancers. *Nutr Rev,* 65**,** S157-66.

MAENO, M., YAMAMOTO, N. & TAKANO, T. (1996) Identification of an antihypertensive peptide from casein hydrolysate produced by a proteinase from Lactobacillus helveticus CP790. *J Dairy Sci,* 79**,** 1316-21.

MARENCO, L., WANG, T. Y., SHEPHERD, G., MILLER, P. L. & NADKARNI, P. (2004) QIS: A framework for biomedical database federation. *J Am Med Inform Assoc,* 11**,** 523-34.

MARTIN, P., FERRANTI, P., LEROUX, C. & ADDEO, F. (2003) *Non-bovine caseins:quantitaive variablity and molecular diversity*.

MARTIN, P., SZYMANOWSKA, M., ZWIERZCHOWSKI, L. & LEROUX, C. (2002) The impact of genetic polymorphisms on the protein composition of ruminant milks. *Reprod Nutr Dev,* 42**,** 433-59.

MATHER, I. H. (2000) A review and proposed nomenclature for major proteins of the milk-fat globule membrane. *J Dairy Sci,* 83**,** 203-47.

MCFADDEN, T. B., AKERS, R. M. & KAZMER, G. W. (1987) Alpha-lactalbumin in bovine serum: relationships with udder development and function. *J Dairy Sci,* 70**,** 259-64.

MEISEL, H. (1997) Biochemical properties of bioactive peptides derived from milk proteins: Potential nutraceuticals for food and pharmaceutical applications. *Livestock Production Science,* 50**,** 125 - 138.

MICKLEM, G., SMITH, R. & K., R. (2006) InterMine. 8.0 ed. Cambridge, U.K., University of Cambridge.

MIKA, S. & ROST, B. (2004) NLProt: extracting protein names and sequences from papers. *Nucleic Acids Res,* 32**,** W634-7.

MOLLOY, M. P., HERBERT, B. R., YAN, J. X., WILLIAMS, K. L. & GOOLEY, A. A. (1997) Identification of wallaby milk whey proteins separated by two-dimensional electrophoresis, using amino acid analysis and sequence tagging. *Electrophoresis,* 18**,** 1073-8.

MURAKAMI, K., LAGARDE, M. & YUKI, Y. (1998) Identification of minor proteins of human colostrum and mature milk by two-dimensional electrophoresis. *Electrophoresis,* 19**,** 2521-7.

NARAYANASAMY, V., MUKHOPADHYAY, S., PALAKAL, M. & POTTER, D. A. (2004) TransMiner: mining transitive associations among biological objects from text. *J Biomed Sci,* 11**,** 864-73.

ODDY, W. H. (2002) The impact of breastmilk on infant and child health. *Breastfeed Rev,* 10**,** 5-18.

ODDY, W. H. (2004) A review of the effects of breastfeeding on respiratory infections, atopy, and childhood asthma. *J Asthma,* 41**,** 605-21.

OINN, T., ADDIS, M., FERRIS, J., MARVIN, D., SENGER, M., GREENWOOD, M., CARVER, T., GLOVER, K., POCOCK, M. R., WIPAT, A. & LI, P. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics,* 20**,** 3045-54.

OKAZAKI, N. & ANANIADOU, S. (2006) Building an abbreviation dictionary using a term recognition approach. *Bioinformatics,* 22**,** 3089-95.

ONO, T., HISHIGAKI, H., TANIGAMI, A. & TAKAGI, T. (2001) Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics,* 17**,** 155-61.

PEREZ, M. D. & CALVO, M. (1995) Interaction of beta-lactoglobulin with retinol and fatty acids and its role as a possible biological function for this protein: a review. *J Dairy Sci,* 78**,** 978-88.

PEREZ, M. D., DIAZ DE VILLEGAS, C., SANCHEZ, L., ARANDA, P., ENA, J. M. & CALVO, M. (1989) Interaction of fatty acids with beta-lactoglobulin and albumin from ruminant milk. *J Biochem (Tokyo),* 106**,** 1094-7.

PEREZ, M. D., SANCHEZ, L., ARANDA, P., ENA, J. M., ORIA, R. & CALVO, M. (1992) Effect of beta-lactoglobulin on the activity of pregastric lipase. A possible role for this protein in ruminant milk. *Biochim Biophys Acta,* 1123**,** 151-5.

PRATT, W. & YETISGEN-YILDIZ, M. (2003) LitLinker: Capturing Connections Across the Biomedical Literature. *K-CAP 2003*.

REBHOLZ-SCHUHMANN, D. (2005) EBIMed.

REBHOLZ-SCHUHMANN, D., KIRSCH, H., ARREGUI, M., GAUDAN, S., RYNBEEK, M. & STOEHR, P. (2006) Protein annotation by EBIMed. *Nat Biotechnol,* 24**,** 902-3.

REICH, C. M. & ARNOULD, J. P. (2007) Evolution of Pinnipedia lactation strategies: a potential role for alpha-lactalbumin? *Biol Lett,* 3**,** 546-9.

SAITO, T. (2008) Antihypertensive peptides derived from bovine casein and whey proteins. *Adv Exp Med Biol,* 606**,** 295-317.

SANO, J., OHKI, K., HIGUCHI, T., AIHARA, K., MIZUNO, S., KAJIMOTO, O., NAKAGAWA, S., KAJIMOTO, Y. & NAKAMURA, Y. (2005) Effect of casein hydrolysate, prepared with protease derived from Aspergillus oryzae, on subjects with high-normal blood pressure or mild hypertension. *J Med Food,* 8**,** 423-30.

SCHREZENMEIR, J. & JAGLA, A. (2000) Milk and diabetes. *J Am Coll Nutr,* 19**,** 176S-190S.

SELO, I., CLEMENT, G., BERNARD, H., CHATEL, J., CREMINON, C., PELTRE, G. & WAL, J. (1999) Allergy to bovine beta-lactoglobulin: specificity of human IgE to tryptic peptides. *Clin Exp Allergy,* 29**,** 1055-63.

SETTLES, B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics,* 21**,** 3191-2.

SHAH, P. K., PEREZ-IRATXETA, C., BORK, P. & ANDRADE, M. A. (2003) Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics,* 4**,** 20.

SHAH, S. P., HUANG, Y., XU, T., YUEN, M. M., LING, J. & OUELLETTE, B. F. (2005) Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics,* 6**,** 34.

SINCLAIR, G. & WEBBER, B. (2004) Classification from Full Text: A Comparison of Canonical Sections of Scientific Papers. *Coling-2004.* Geneva.

SMALHEISER, N. R. (2005) The arrowsmith project: 2005 status report. *Discovery Science, Proceedings.*

SMALHEISER, N. R. & SWANSON, D. R. (1996a) Indomethacin and Alzheimer's disease. *Neurology,* 46**,** 583.

SMALHEISER, N. R. & SWANSON, D. R. (1996b) Linking estrogen to Alzheimer's disease: an informatics approach. *Neurology,* 47**,** 809-10.

SMALHEISER, N. R. & SWANSON, D. R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed,* 57**,** 149-53.

SMITH, L., RINDFLESCH, T. & WILBUR, W. J. (2004) MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics,* 20**,** 2320-1.

SPITSBERG, V. L. (2005) Invited review: Bovine milk fat globule membrane as a potential nutraceutical. *J Dairy Sci,* 88**,** 2289-94.

SRINIVASAN, P. & LIBBUS, B. (2004) Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics,* 20 Suppl 1**,** I290-I296.

STAPELFELDT, H., PETERSEN, P. H., KRISTIANSEN, K. R., QVIST, K. B. & SKIBSTED, L. H. (1996) Effect of high hydrostatic pressure on the enzymic hydrolysis of beta-lactoglobulin B by trypsin, thermolysin and pepsin. *J Dairy Res,* 63**,** 111-8.

STEIN, T., MORRIS, J. S., DAVIES, C. R., WEBER-HALL, S. J., DUFFY, M. A., HEATH, V. J., BELL, A. K., FERRIER, R. K., SANDILANDS, G. P. & GUSTERSON, B. A. (2004) Involution of the mouse mammary gland is associated with an immune cascade and an acute-phase response, involving LBP, CD14 and STAT3. *Breast Cancer Res,* 6**,** R75-91.

STEIN, T., PRICE, K. N., MORRIS, J. S., HEATH, V. J., FERRIER, R. K., BELL, A. K., PRINGLE, M. A., VILLADSEN, R., PETERSEN, O. W., SAUTER, G., BRYSON, G., MALLON, E. A. & GUSTERSON, B. A. (2005) Annexin A8 is up-regulated during mouse mammary gland involution and predicts poor survival in breast cancer. *Clin Cancer Res,* 11**,** 6872-9.

STEPHENS, M., PALAKAL, M., MUKHOPADHYAY, S., RAJE, R. & MOSTAFA, J. (2001) Detecting gene relations from Medline abstracts. *Pac Symp Biocomput***,** 483-95.

SUCHYTA, S. P., SIPKOVSKY, S., HALGREN, R. G., KRUSKA, R., ELFTMAN, M., WEBER-NIELSEN, M., VANDEHAAR, M. J., XIAO, L., TEMPELMAN, R. J. & COUSSENS, P. M. (2003a) Bovine mammary gene expression profiling using a cDNA microarray enhanced for mammary-specific transcripts. *Physiol Genomics,* 16**,** 8-18.

SUCHYTA, S. P., SIPKOVSKY, S., KRUSKA, R., JEFFERS, A., MCNULTY, A., COUSSENS, M. J., TEMPELMAN, R. J., HALGREN, R. G., SAAMA, P. M., BAUMAN, D. E., BOISCLAIR, Y. R., BURTON, J. L., COLLIER, R. J.,

DEPETERS, E. J., FERRIS, T. A., LUCY, M. C., MCGUIRE, M. A., MEDRANO, J. F., OVERTON, T. R., SMITH, T. P., SMITH, G. W., SONSTEGARD, T. S., SPAIN, J. N., SPIERS, D. E., YAO, J. & COUSSENS, P. M. (2003b) Development and testing of a high-density cDNA microarray resource for cattle. *Physiol Genomics,* 15**,** 158-64.

SUN, Z., ZHANG, Z., WANG, X., CADE, R., ELMIR, Z. & FREGLY, M. (2003) Relation of beta-casomorphin to apnea in sudden infant death syndrome. *Peptides,* 24**,** 937-43.

SWAISGOOD, H. E. (2003) *Chemistry of the caseins*, Kluwer academic.

SWANSON, D. R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med,* 30**,** 7-18.

SWANSON, D. R. (1988) Migraine and magnesium: Eleven neglected connections. *Perspect Biol Med,* 34**,** 526-557.

SWANSON, D. R. (1990) Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspect Biol Med,* 33**,** 157-86.

SWANSON, D. R. (2006) Atrial fibrillation in athletes: Implicit literature-based connections suggest that overtraining and subsequent inflammation may be a contributory mechanism. *Medical Hypotheses,* 66**,** 1085-1092.

SWANSON, D. R., SMALHEISER, N. R. & BOOKSTEIN, A. (2001) Information discovery from complementary literatures: Categorizing viruses as potential weapons. *Journal Of The American Society For Information Science And Technology,* 52**,** 797-812.

SWANSON, D. R., SMALHEISER, N. R. & TORVIK, V. I. (2006) Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal Of The American Society For Information Science And Technology,* 57**,** 1427-1439.

TANABE, L. & WILBUR, W. J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics,* 18**,** 1124-32.

TRUSWELL, A. S. (2005) The A2 milk case: a critical review. *Eur J Clin Nutr,* 59**,** 623-31.

VAN SANTVOORT, H. C., BESSELINK, M. G., VAN MINNEN, L. P., TIMMERMAN, H. M., AKKERMANS, L. M. & GOOSZEN, H. G. (2006) [Potential role for probiotics in the prevention of infectious complications during acute pancreatitis]. *Ned Tijdschr Geneeskd,* 150**,** 535-40.

VERHEUGT, F. W. (2004) Is antithrombiotc therapy a risk-free and beneficial treatment for patients with heart failure? *Nature Clinical Practice,* 1**,** 80 - 81.

WARD, R. E. & GERMAN, J. B. (2004) Understanding milk's bioactive components: a goal for the genomics toolbox. *J Nutr,* 134**,** 962S-7S.

WATERFIELD, B. (2008) 'Friendly bacteria' products linked to 24 deaths. *The Telegraph.*

WEEBER, M., KLEIN, H., ARONSON, A. R., MORK, J. G., DE JONG-VAN DEN BERG, L. T. & VOS, R. (2000) Text-based discovery in biomedicine: the architecture of the DAD-system. *Proc AMIA Symp***,** 903-7.

WEEBER, M., KLEIN, H., DE JONG-VAN DEN BERG, L. T. W. & VOS, R. (2001) Using concepts in literature-based discovery: Simulating Swanson's Raynaud-

fish oil and migraine-magnesium discoveries. *Journal Of The American Society For Information Science And Technology,* 52**,** 548-557.

WEEBER, M., VOS, R., KLEIN, H., DE JONG-VAN DEN BERG, L. T., ARONSON, A. R. & MOLEMA, G. (2003a) Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J Am Med Inform Assoc,* 10**,** 252-9.

WEEBER, M., VOS, R., KLEIN, H., DE JONG-VAN DEN BERG, L. T. W., ARONSON, A. R. & MOLEMA, G. (2003b) Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide. *Journal Of The American Medical Informatics Association,* 10**,** 252-259.

WHITE, J. M., WILLIAMS, G., SAMOUR, J. H., DRURY, P. J. J. & CHEESEMAN, P. (1985) The Composition Of Milk From Captive Aardvark (Orycteropus-Afer). *Zoo Biology,* 4**,** 245-251.

WOLD, A. E. & ADLERBERTH, I. (2000) Breast feeding and the intestinal microflora of the infant--implications for protection against infectious diseases. *Adv Exp Med Biol,* 478**,** 77-93.

WREN, J. D. (2004) Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics,* 5**,** 145.

WROBLEWSKA, B., KARAMAC, M., AMAROWICZ, R., SZYMKIEWICZ, A., TROSZYNSKA, A. & KUBICKA, E. (2004) Immunoreactive properties of peptide fractions of cow whey milk proteins after enzymatic hydrolysis. *International Journal Of Food Science And Technology,* 39**,** 839-850.

YAMADA, M., MURAKAMI, K., WALLINGFORD, J. C. & YUKI, Y. (2002) Identification of low-abundance proteins of bovine colostral and mature milk using two-dimensional electrophoresis followed by microsequencing and mass spectrometry. *Electrophoresis,* 23**,** 1153-60.

YEH, A. S., HIRSCHMAN, L. & MORGAN, A. A. (2003) Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics,* 19 Suppl 1**,** i331-9.