Chapter 16

IDENTIFICATION OF BIOLOGICAL RELATIONSHIPS FROM TEXT DOCUMENTS

Mathew Palakal, Snehasis Mukhopadhyay, and Matthew Stephens

Indiana University Purdue University Indianapolis, Indianapolis, IN 46202

Chapter Overview

Identification of relationships among different biological entities, e.g., genes, proteins, diseases, drugs and chemicals, etc, is an important problem for biological researchers. While such information can be extracted from different types of biological data (e.g., gene and protein sequences, protein structures), a significant source of such knowledge is the biological textual research literature which is increasingly being made available as large-scale public-domain electronic databases (e.g., the Medline database). Automated extraction of such relationships (e.g., gene A inhibits protein B) from textual data can significantly enhance biological research productivity by keeping researchers up-to-date with the state-of-the-art in their research domain, by helping them visualize biological pathways, and by generating likely new hypotheses concerning novel interactions some of which can be good candidates for further biological research and validation. In this chapter, we describe the computational problems and their solutions in such automated extraction of relationships, and present some recent advances made in this area.

Keywords

biological objects; associations; text mining; transitivity; flat relationships; directional relationships; hierarchical relationships

1. INTRODUCTION

The scientific literature is an important source of knowledge for the scientist during the course of study of any research problem. The huge and rapidly increasing volume of scientific literature makes finding relevant information increasingly difficult. Content level information rather than collection level information is needed for scientific research. The availability of scientific literature in electronic format, such as Medline, has made the development of automated text mining systems and hence "data-driven discovery" possible. Text mining enables analysis of large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns or knowledge (Tan, 1999). Text mining has a very high potential for knowledge discovery, as the most natural form of reporting, storing, and communicating information is text. Informatics tools can assist the traditional hypothesis-driven research (Smalheiser, 2001). Many hypotheses are formed by extrapolating the current knowledge; for example, if we know that apoptosis in breast cancer is mediated by calpain, we can ask if apoptosis in other related cancer (e.g., ovarian cancer) is also mediated by calpain.

The query "breast cancer" on Medline returned 128,171 documents on June 21, 2004. This shows the large volume of scientific information available in the form of text. It is impossible or impractical for anyone to read through all of these documents to find the relevant information. It is even more difficult to capture the knowledge in those documents. Researchers and scientists are challenged by this increasing knowledge gap. Associations among biological objects such as genes, proteins, molecules, processes, diseases, drugs and chemicals, are one such form of underlying knowledge. For example, Swanson (Swanson and Smalheiser, 1997) found an association between magnesium and migraine headaches that was *not explicitly reported in any one article*, but based on *associations extracted from different journal titles*, and later validated experimentally.

In this chapter, we describe the progress made in the development of a complete "knowledge base" of associations among biological objects, such as those mentioned above, that are important for biologists to study and understand specific biological processes. The term object refers to any biological object (e.g. protein, gene, cell cycle, etc.) and relationship refers to an action one object has on another. Biological relationships discovered from literature and experiments can be used to set up templates for biologists to model a biological process and to formulate new hypotheses for guided laboratory research.

The main goals of this chapter are to describe the progress made in: (a) developing a very large knowledge base, called BioMap, using the entire

Medline collection (over 14 million) of literature documents, and (b) developing an interactive knowledge network for users to access this secondary knowledge (BioMap) along with its primary databases such as Medline, GenBank, etc., in an integrated manner based on a specific area of problem enquiry. In order to build the BioMap and its associated access "window" (the knowledge network), various algorithms and tools need to be developed for: (i) identifying biological object names; (ii) discovering object-object relationships; (iii) creation of the knowledge base (BioMap); (iv) a hypergraph realization of the knowledge network (generating pathways and hypothesis) in response to a user query, and, (v) global access capability for the entire system.

Identification of biological objects and their relationships from free running text is a very difficult problem. This problem is compounded by several factors, specifically, when multiple objects and multiple relationships need to be detected. Typically, the extraction of object relationships involves object name identification, reference resolution, ontology and synonym discovery, and finally extracting object-object relationships. We describe a multi-level hybrid approach that incorporates statistical, connectionist, and N-Gram models along with multiple dictionaries to handle the multi-object identification and relationship extraction problem for BioMap.

The relationships thus discovered from the entire Medline collection are to be maintained in a relational database along with the specific links to literature sources, genes and protein sequence databases. A user can access this knowledge base using simple or complex queries, such as a disease name, a set of gene names, or any such combinations. Unlike in traditional databases, the outcome of a user query will be a complex set of data with multiple associations among them. Hence, the results of a query will be constructed as a *knowledge network* (knowledge view of BioMap) and presented to the user.

The knowledge network is to be constructed as a hypergraph "on the fly" based on each user query. A hypergraph is an extension of a graph in the sense that each hyperedge can connect more than two vertices, thus allowing to connect relationships among multiple objects simultaneously. A system based on such a hypergraph model has a number of advantages: the model is independent from updates to the underlying databases; it enables the formation of hypergraphs from entities in different databases; it allows the system to be accessed by multiple users simultaneously, each with an independent hypergraph; further queries can be made to a hypergraph to obtain a better "focused" view of the knowledge base; it reduces the need to access the knowledge base multiple times when a hypergraph is to be shown to a remote user; and most importantly, the interaction with the user can be made faster when a query is made on the hypergraph since it is available in the local computer memory. Furthermore, the edges and vertices of the hypergraph are "live" (made as hyperlinks) allowing the user to access primary data sources and bioinformatics tools pertinent to the information contained in the knowledge network.

The key innovative features of the proposed BioMap are that it is adaptable and scalable, and that the core knowledge base will be constructed from a very large collection of documents (Medline) to make the system robust. The adaptation feature requires that the system should have the ability to learn new problem domains without having to rebuild the system as in the case of fully rule-based or grammar-based approaches. The scalability feature allows the system to continue to develop its knowledge base as new information arrives in the literature databases or information is incorporated from other data sources (e.g. *Science*, *Nature*, etc.).

2. OVERVIEW OF THE FIELD

2.1 Background

Ever since the emergence of the field of Bioinformatics, it has been of great interest for both the informatics and the biology communities to develop automatic methods to extract embedded knowledge from literature data. Dealing with literature data in free running text is a challenging problem that has been well studied by natural language processing (NLP) and artificial intelligence (AI) communities with some success. For example, several works report relationship extraction among biological objects (Ono et al., 2001; Humphreys et al., 2000; Thomas et al., 2000; Proux et al., 2000; Marcotte et al., 2001; Oyama et al., 2002) and biological object recognition problems (Leroy and Chen, 2002; Tanabe and Wilbur, 2002; Krauthammer et al., 2000). Some work has also been reported on document clustering (Iliopoulus et al., 2001; Nobata et al., 2000) and pathway identification (Sanchez et al., 1999; Park et al., 2001; Ng and Wong, 1999). Current progress in supporting biomedical research activities through published literature can be broadly classified into two categories: (i) Biological information extraction (IE), and (ii) Development of "tools of the trade" bioinformatics tools.

2.2 Biological Information Extraction

Mining of literature databases to discover information relevant to biological relationships and pathways involves two key tasks: first, identification of biological object names, and second, identification of relationships among these objects. Several works report research on the identification of biological objects. The most successful tagging system, described in (Fukuda et al., 1998), is a rule-based system, called PROPER, that was specifically designed to extract protein names from the text using proper noun dictionaries and a pattern dictionary. The results of the PROPER tagging method were evaluated using precision and recall and yielded an accuracy of 98.84% and 94.70% respectively. These results, however, do not include distinction between gene and protein names and leave out words that may not be in the target object list. Also, it was designed to tag only one object (i.e., proteins), which is not adequate for extracting different object names.

Collier (Collier et al., 2000a) proposed a stochastic approach to tagging biological objects. Their model utilized 100 Medline abstracts using a pre-specified annotation method and used this data to train a Hidden Markov Model to tag similar objects. The results of this method, given as F-scores, combine recall and precision (Chinchor, 1995). In the case of tagging the proteins, the best F-score reported was 0.759 using the one hundred hand-tagged abstracts. For DNA it was significantly less at 0.472. The highest average F-score calculated was 0.728. It was assumed that the increased training data would improve the performance but how much more training data would be necessary to achieve performance above the desired F-score of 0.9 was not reported.

In order to overcome the limitation of hand tagging, Hatzivassilou (Hatzivassiloglou et al., 2001) trained models using an extensive dictionary of unambiguous gene terms from the GeneBank database. Using a nine million-word corpus, they managed to distinguish between three biological entities using a Bayesian classifier with accuracy of 80%. A two-way classifier jumps up to an accuracy of 85%. The fact that accuracy starts to decline with the addition of more classes indicates that such a method would not scale up when tagging multiple object types. In conclusion, the current methods for tagging biological objects fall short because they either cannot tag more than one object (Fukuda et al., 1998) or they rely heavily on hand-tagged training data (Collier et al., 2000a).

Natural language-based parsers have also been used on biological literature to extract relationships such as protein-protein interactions. For example, a full parser was used in (Yakushiji et al., 2000) to extract information from biomedical papers. One reported experiment consisted of 179 sentences from an annotated corpus of Medline abstracts. The first 97 sentences were used to determine the accuracy of the system. Out of 133 argument structures in those 97 sentences, 23% were extracted uniquely, 24% with ambiguity, and 53% were not extracted. Another NLP system reported in (Friedman et al., 2001) is called Genie, consisting of a term tagger, preprocessor and parser. The term tagger uses BLAST (Atschul et al., 1990) techniques, specialized rules, and external knowledge sources to identify and tag genes and proteins in the text articles. The Genie system had a measured sensitivity of 54% and a specificity of 96%. Although the reported accuracy of the system is good, it does not take into consideration where the interaction takes place and under what conditions.

A method to identify Gene-pair relationships from a large collection of text documents is reported in (Stephens et al., 2001). The goal is to discover pairs of genes from a collection of retrieved text documents such that the genes in each pair are related to one other in some manner. Details of this method are discussed in Section 3b(iv) under title "rFinder-I." The results of this study indicated that finding the actual nature of the relationship between proteins had a specificity of 67% in the unknown pathway and specificity of 50% in the known pathway. The potential drawback of this approach is that it finds only gene-pair relationships; relationships that occur indirectly across sentences will not be found. Another study (Craven et al., 1999) proposed a learning method to extract relationships and organize these relationships as structured representations or knowledge bases. This study, primarily focused on protein related interactions, reports 77% precision and 30% recall on a corpus of 633 sentences. EDGAR is another natural language processing system (Rindflesch et al., 2000) that extracts relationships between cancer-related drugs and genes from biomedical literature. Again, the scope is limited to few biological objects and their relationships.

Recently, support vector machine (SVM) based approaches (Kazama et al., 2002; Steffen et al., 2003) showed promising results for biological entity identification. Using SVM, a named entity task is formulated as the classification of each word with context to one of the classes that represent the region information and entity's semantic class. The best results reported so far show a precision of 71.4% and a recall of 72.8%, corresponding to an F-measure of 72.1%, for the closed division for only gene name identification.

In summary, the current methods for tagging biological objects fall short because they either cannot tag multiple objects or they rely heavily on hand-tagged training data. The techniques that are used for extracting relationships using NLP are too specific to be extended to new domains without creating a large number of new rules for new relationships. Also, most importantly, current approaches do not look into the creation of a "knowledge base" of all possible relationships for biological problem domain or domains, so that the research community can not only retrieve relevant literature but also retrieve and view the embedded knowledge.

2.3 Bioinformatics Tools

Numerous bioinformatics tools also exist that closely or loosely connect to provide literature support for biomedical research. These systems have relied on annotation of the biomedical literature, with the most successful system being the Online Mendelian Inheritance in Man (OMIM) database and its associated morbid map (OMIM). While OMIM has been very successful in the development of an annotated disease-based database, the very nature of its annotation means *OMIM only presents the well established and proven associations for a given disease*. As such, the OMIM database is not capable of finding novel associations with respect to different diseases of interest. For example, for the discovery of novel gene-disease relationships one needs a list of all the possible gene-disease relationships, even if currently unproven, such that a scientist may find a weak gene-disease relationship that is strengthened by the addition of their own research data.

Databases have been designed and tools built to search the biomedical literature as well. The best is PubMed, the searchable database related to biomedical literature present in Medline. This database, with over 14 million references, is the most comprehensive listing of biomedical literature in the world. These tools can be used to download and parse the appropriate data to a secondary database that can be examined based on the users needs. Secondary databases that perform these functions include MedMiner, which allows one to query Genecards using terms related to physiologic pathways and receive back a list of genes involved in that pathway. In addition, gene or drug names can be sent to PubMed to identify the biomedical literature by searching the abstracts, Keywords, and MeSH terms. A useful function that is not present in MedMiner is the capability to comprehensively search for all genes related to a keyword. Thus MedMiner does not allow the desired degree of flexibility in user search terms and the comprehensive search of all key biological names.

PubGene (Jenssen et al., 2001) uses a similar design to allow the user to query genes using the HUGO approved gene symbols in its database. This database contains relationships identified through searches of Medline and identifies pairs of genes that are mentioned in the same abstract or correlated by GO (Ashburner, 2000) classifiers. The PubGene query system returns a graphical representation of the gene-gene relationships mentioned in the same reference as the queried gene. However, *the focus of the PubGene process is to identify gene-gene relationships and not gene-search term relationships* ("search term" can be gene, protein, drug, etc.).

In summary, the available databases offer useful information related to genes and their interrelationships in the biomedical literature, however, there is a lack of a truly flexible user-driven data mining system for multiple biological objects, even for all genes. More importantly, most of the existing tools *provide well established and proven associations* in actively investigated areas and they do not provide information on associations that are less organized and obvious. The proposed BioMap and its associated tools described herein are designed to specifically address these issues.

3. CASE STUDIES

Different types of biological relationships can be extracted from literature documents. These include flat relationships, directional relationships, and hierarchical relationships. Flat relationships simply state there exists a relationship between two biological entities. Directional relations also indicate the direction of the relationship that actually applies, for example, "A inhibits B" or "A is inhibited by B." In this section, we present three case studies in biological association discoveries. The first two studies (described in section 3.1 and 3.2) illustrate in a comprehensive way all the problems arising in biological relationship finding and some computational approaches to their solutions. The second study deals with an important extension of the basic association discovery methods, i.e., using transitivity property to postulate implicit potentially novel associations. The third study looks into discovering directional and hierarchical associations using text mining approaches.

3.1 Identification of Flat Relationships from Text Documents

In this case study we present a Thesaurus-based text analysis approach to discover the existence and the functional nature of relationships between one single biological object (e.g. gene) relating to a problem domain of interest. The approach relies on multiple Thesauri, representing domain knowledge as gene names and terms describing gene functions. These Thesauri can be constructed using existing organizational sources (e.g., NCBI and EBI), by consulting experts in the domain of interest, or by the users themselves. Thesauri can also be constructed using automated vocabulary discovery techniques being developed by the Information Extraction (IE) or Information Retrieval (IR) communities. In its simplest form, a Thesaurus consists of a linear list of terms and associated concepts. The process involves Thesaurus-based content representation of the retrieved documents, identification of associations (relationships) and finally, detecting gene functionality from the represented retrieved document set. These primary steps are described in detail in the following sections along with some experimental results.

3.1.1 Text Document Representation

The document representation step converts text documents into structures that can be efficiently processed without the loss of vital content. At the core of this process is a thesaurus, an array T of atomic tokens (e.g., a single term) each identified by a unique numeric identifier culled from authoritative sources or automatically discovered. A thesaurus is an extremely valuable component in term-normalization tasks and for replacing an uncontrolled vocabulary set with a controlled set (Rothblatt et al., 1994). Beyond the use of the thesaurus, the *tf.idf* (the term frequency multiplied with inverse document frequency) algorithm (Rothblatt et al., 1994) is applied as an additional measure for achieving more accurate and refined discrimination at the term representation level. In this formula, the *idf* component acts as a weighting factor by taking into account inter-document term distribution, over the complete collection given by:

$$W_{ik} = T_{ik} \times \log(N / n_k) \qquad (1)$$

Where T_{ik} is the number of occurrences of term T_{k} in document i, $I_k = \log(N/n_k)$ is the inverse document frequency of term T_{k} in the document base, N is the total number of documents in the document base, and n_{k} is the number of documents in the base that contain the given term T_{k}.

As document representation is conducted on a continuous stream, the number of documents present in the stream may be too few for the *idf* component to be usefully applied. To deal with this, a table is maintained containing total frequencies of all thesaurus terms in a sufficiently representative collection of documents as a base (randomly sampled documents from the source used as the training set). It is worth pointing out that such a table can be pre-constructed off-line before any on-line analysis of retrieved documents is attempted. The purpose of the document representation step is to convert each document to a weight vector whose dimension is the same as the number of terms in the thesaurus and whose elements are given by the above equation.

3.1.2 Gene-pair Relationship

The goal here is to discover pairs of genes from a collection of retrieved text documents such that the genes in each pair are related to one other in some manner. Whether two genes are to be related depends on somewhat subjective notion of "being related." We have investigated Gene-pair discovery from a collection of Medline abstracts using the Vector-Space *tf**idf* method and a thesaurus consisting of Gene terms. Each Gene term, in turn, contains several synonymous keywords that are gene names. Each document d_i is converted to a M dimensional vector W_i where W_{ik} denotes the weight of the k^{th} gene term in the document and M indicates the number of terms in a Thesaurus. W_{ik} is computed by equation 1 described in Section 3.1.1.

It is clear that W_{ik} increases with term frequency T_{ik}. However, it decreases with n_k, i.e., if a gene term occurs in increasingly larger number of documents in the collection, it is treated as a common term and its weight is decreased.

Once the vector representation of all documents are computed, the association between two gene terms k and l is computed as follows:

$$association[k][l] = \sum_{i=1}^{N} W_{ik} * W_{il} \quad \text{k} = 1...\text{m}, l = 1...\text{m} \qquad (2)$$

For any pair of gene terms co-occurring in even a single document, the *association*[k][l] will be non-zero and positive. However, the relative values of *association*[k][l] will indicate the product of the importance of the k^{th} and l^{th} term in each document, summed over all documents. This computed association value is used as a measure of the degree of relationship between the k^{th} and l^{th} gene terms. A decision can be made about the existence of a strong relationship between genes using a user-defined threshold on the elements of the Association matrix.

3.1.3 Functional Nature of Relationships Between Gene-pairs

Once a "relationship" has been found between genes, the next step is to find out what that relationship is. This requires an additional thesaurus containing terms relating to possible relationships between genes that a user may be interested in. This thesaurus is then applied to sentences, which contain co-occurring gene names. If a word in the sentence containing co-occurrences of genes matches a relationship in the thesaurus, it is counted as a score of one. The highest score over all sentences for a given relationship is then taken to be the relationship between the two genes or proteins. A score of as little as one could be significant because a relationship may be only mentioned in one abstract. A higher score, however, would be more likely to indicate that relationship because they are often reiterated in multiple abstracts. The following equation summarizes the relationship:

$$score[k][l][m] = \sum_{i=1}^{S} p_i \, ; (p_i = 1 : Gene_k, Gene_l, \text{Re}lation_m \text{ all occur in sentence} \, i) \quad (3)$$

where, S is the number of sentences in the retrieved document collection, p_i is a score equal to 1 or 0 depending on whether or not all terms are present, and Gene_k refers to the gene in the gene thesaurus with index *k*, and relation_m refers to the term in the relationship thesaurus with index *m*. The functional nature of the relationship is chosen as $arg_{\max}$ score[k][l][m].

The idea is to narrow down the search to a few relationships which the user can check. If a functional relationship cannot be found the user can still check against articles where the terms co-occurred to see if a function might have been missing from the function thesaurus containing the relationships. Overall, this will help the user to quickly develop potential pathways and speed up the process of finding genetic interactions.

3.1.4 Experimental Results

Two experiments show how this technique performs in accuracy and as a tool for discovering a legitimate pathway based on retrieved data. The list of potential relations used for both examples, determined manually using a Molecular Biology text book (Salton, 1989), is shown in Table 16-1. The first experiment uses the gene list shown in Table 16-2.

Table 16-1. The Thesaurus of Relationships

"activates, activator"	"inhibits, inhibitor"	"phosphorylates"		
"binds, binding, complexes"	"catalyst, catalyses"	"hydrolysis, hydrolyzes"		
"cleaves"	"adhesion"	"donates"	"regulates"	"induces"
"creates"	"becomes"	"transports"	"exports"	"releases"
"suppresses, suppressors"				

This list includes genes and proteins not taken from any particular pathway but is associated with cell structure and muscle cells.

Table 16-2. Thesaurus of Genes (Unknown Pathway)

"actinin"	"actn2"	"ank1, ankyrin"	"atf4"	"ca3"	"CD36"	"cd54"
"COI"	"cox1"	"CSE1"	"cst3"	"desmin"	"FKBP51"	"FKBP54"
"FUS, TLS"	"GAPDH"	"hmsh2"	"hrv"	"hsp90"	"importin"	
"lim"	"mcm4"	"myoglobin"	"nebulin"	"nfatc"	"myosin"	
"nop-30"	"NPI-1"	"p55"	"titin"	"ubiquinone"	"filamin"	

The training documents are created by taking an equal number of abstracts from the Medline database for each gene. Altogether, 5,072 abstracts were used.

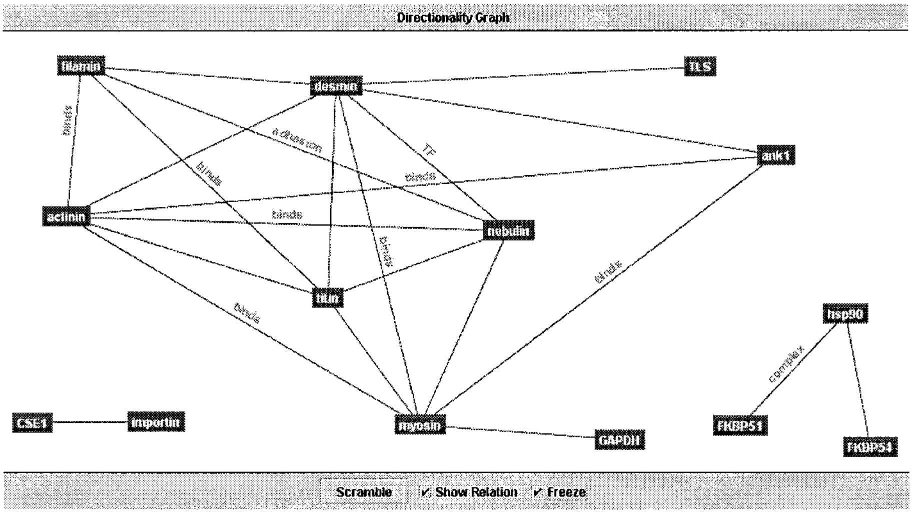


Figure 16-1. Graph showing relationships between genes in Known Pathway. The higher the Association strength the closer the genes appear on the graph. In this way the related genes are clustered together and can be picked out.

A graphical presentation of the unknown pathway (Table 16-2) is shown in Figure 16-1. The relationship discovery aspect of this method was excellent. This was verified by looking at the actual abstracts on the basis of which associations were computed. The strong central cluster includes proteins involved in construction of the cytoskeleton. The cluster containing CSE1 and *importin* are involved in the process of recycling *importin* and the other cluster contains proteins involved in making a steroid receptor complex. More details about the results and discussions can be found in (Stephens et al., 2001).

3.2 TransMiner: Formulating Novel, Implicit Associations Through Transitive Closure

An important question in biological knowledge management is whether it is possible to generate novel hypotheses concerning associations between biological objects, based on existing associations as presented in the literature. We have developed a system called TransMiner, which aims to identify transitive associations by using graph theoretic properties, in particular the transitivity property, on an underlying association graph. A strong motivation for the use of such transitivity property was provided by Swanson (Swanson and Smalheiser, 1997) and his co-workers. The idea is that if, according to existing literature, object A is related to object B, and object B is related to object C, then there is a likelihood of A being related to C, even though the last association may not have been explicitly reported. Moreover, considering such transitive (implicit) association between A and C, the likelihood of its existence is increased if more and more intermediate objects (object B) are found in the literature. Further, such transitive property can be extended through any number of intermediate nodes, as incorporated in the transitive closure of the original graph. Swanson developed a system called ARROWSMITH (Swanson and Smalheiser, 1997) that automated the one-step transitive relationship discovery process by considering only document titles and one intermediate node. TransMiner generalizes it by considering entire document abstracts (also full-text articles, if available) in addition to document titles, and also extending the transitivity property to the complete transitive closure of the original graph.

Swanson made seven medical discoveries by analyzing medical literature and applying the one-step transitivity property on titles (including the famous prediction of the magnesim-migraine association, before it was biologically verified). Smalheiser (Smalheiser, 2002) a collaborator of Swanson used ARROWSMITH to discover that genetic packaging technologies such as DEAE-dextran, cationic liposomes and cyclodextrins are plausible candidates to enhance infections caused by viruses delivered via an aerosol route – despite the fact that no studies had been reported that examined this issue directly.

Another novel feature of TransMiner is an iterative retrieval and association extraction process in an attempt to verify potential new associations from literature in an effort to overcome limited initial document set size. This prevents processing an inordinately large document set unnecessarily (possibly the entire MedLine!).

3.2.1 Transitive Association Discovery – Methods and Techniques

Relations are ways in which things can stand with regard to one another or to themselves (Honderich, 1995). Relation R is transitive if R (x, y) and R (y, z) imply R (x, z). In symbols, R is transitive if and only if $\forall x \forall y \forall z$ (($Rxy \wedge Ryz$) Rxz).

Transitive Closure:

The transitive closure of a graph G is the graph G* such that there is an edge from vertex A to vertex C in G* if there is a path from A to C in G. The traditional Warshall's (Warshall, 1962) algorithm can be used to compute the transitive closure of the association graph. Given a directed graph G = (V, E) where, V is the set of vertices and E is the set of edges, represented by an adjacency matrix A[i,j], where A[i,j] = 1 if (i,j) is in E, compute the matrix P, where P[i,j] is 1 if there is a path of length greater than or equal to 1 from i to j. Thus, defining $A^0 = I$ (the Identity matrix), and $A^i = A^{i-1}.A$ for all i, where the matrix multiplication is Boolean,

$$P = \sum_{i=1}^{\infty} A^i \tag{4}$$

This algorithm extends paths by joining existing paths together. The transitive closure of a symmetric matrix (undirected graph) is also a symmetric matrix (undirected graph).

Mining Direct and Transitive Associations from Potential Transitive Associations:

The newly discovered potential transitive associations must be checked to see if those associations are indeed 'direct' (explicitly found in any of the Medline documents). We used an automated way (Algorithm 1) to find those associations that are direct and that are transitive, by submitting the two nodes (objects) of a potential transitive relationship to the Medline database with 'AND' operator in the query iteratively for all potential transitive object pairs. The documents will be retrieved only if both the objects are present in the document. For any pair of objects representing a potential transitive relationship, if the document set retrieved is non-zero, then by the principle of co-occurrence we can conclude that there exists a possibility of association between this object pair and that the association is direct. The association strength of these newly discovered 'direct' associations are given by the product of tf.idf weight of both nodes (objects) summed over all the documents retrieved for the object pair. The rest of the potential transitive associations with zero strength are implicit or transitive. These transitive associations are candidates for hypothesis generation. For these transitive associations there are no documents in Medline at present that have both the objects in their contents.

Algorithm 1: Transitive Association Discovery

1. Potential transitive associations are the difference between the transitive closure (G*) and the initial association graph (G).
2. Find the object pair for each potential transitive association and construct the Medline URL query using 'AND' operator.
3. Retrieve documents for this object pair and calculate the association strength between the object pair.
4. If the association strength is not zero, the association is direct. Keep the object pair in G*

5. If the association strength is zero, the association is transitive. Remove the object pair from G*
6. Repeat steps 2, 3, 4, and 5 for all the potential transitive associations discovered to get G' that contains the initial direct associations G and the newly discovered direct associations.

Ranking transitive associations:

Ranking of the transitive associations that are new and potentially meaningful associations will help the user to select associations (hypotheses) that can be further investigated in detail. Transitive association strength cannot be calculated directly as done in the case of direct associations, as there is no co-occurrence in any document between the nodes "A" and "C" of a transitive association. The transitive association strength is defined as the sum of weight of all words "B" that co-occur with both nodes "A" and "C" of a transitive association (intersection of words that co-occur with A and words that co-occur with B). This is based on the idea that if there is a strong link in the form of A-B-C then the possibility of AC association becoming true is more.

3.2.2 Experimental Results - Association Discovery among Breast Cancer Genes

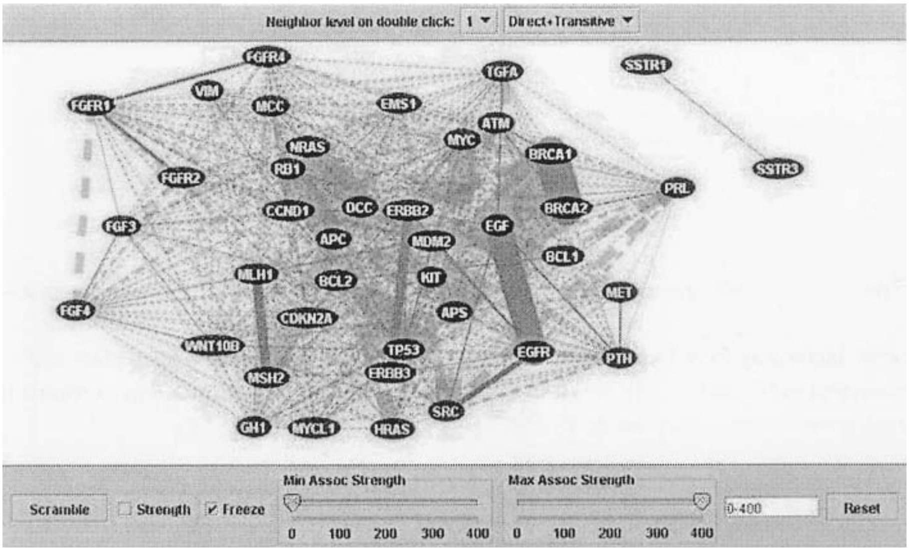
This validation study attempted to use TransMiner to extract gene-gene associations relevant to the disease of Breast Cancer. A list of fifty-six gene symbols related to breast cancer was made from Baylor College of Medicine, Breast Cancer Gene Database (Baasiri et al., 1999) and the GeneCards database (Rebhan et al., 1997). These gene names are given in Table 16-3 and formed the dictionary for the validation study.

Table 16-1. Fifty-six breast cancer genes

APC	APS	ATM	BCL1	BCL2	BRCA1	BRCA2	CCND1
CDKN2A	COL18A1	DCC	EGF	EGFR	EMS1	ERBB2	ERBB3
MSH2	MLH1	FGF3	FGF4	FGFR1	FGFR2	FGFR4	GH1
GRB7	HRAS	IGF1R	KIT	KRAS2	MYCL1	IGF2R	MCC
MDM2	MET	MYC	NF2	NRAS	PGR	PHB	PLAT
PLG	PRL	PTH	PTPN1	RB1	SSTR1	SSTR2	SSTR3
SSTR4	SSTR5	SRC	TGFA	TP53	TSG101	VIM	WNT10B

The initial document set was 5000 Medline documents. The initial association discovery extracted 87 direct associations (i.e., association pairs with non-zero weights). This formed that initial graph G, in which the gene pair BRCA1-BRCA2 was found to have the highest association strength, which is expected. Application of Warshall's transitive closure algorithm on

G (to calculate the transitive closure G* of G) yielded 655 potential new transitive gene pair associations were obtained in (G* - G). The iterative retrieval and validation process identified 296 of them as direct associations (i.e., mentioned explicitly in the literature, although not mentioned in the 5000 original documents) and the remaining 359 as transitive association.


Neighbor level on double click: 1 ▼ Direct+Transitive ▼

□ Strength ☑ Freeze

0-400

Figure 16-2. The initial direct associations among 56 gene symbols based on 5,000 Medline documents (blue edges), the direct associations discovered from potential transitive associations (blue dash edges) based on the presence of non-zero association in Medline database and the transitive associations (pink dash edges)

Figure 16-2 is a color-coded graphical display of all the associations. Based on manual evaluation of the 87 initial gene pair associations discovered by TransMiner, 75 (86.21%) gene pairs were found to have some valid biological association. Similarly, out of 296 direct gene pair associations discovered from potential transitive associations, 237 (80.06%) gene pairs were found to have biological association based on expert evaluation.

The detailed results and evaluations are available at http://sifter.cs.iupui.edu/~sifter/transMiner/TransMinerBCResults.html

3.3 Identification of Directional and Hierarchical Relationships

The association discovery methods described in Sections 3.1 and 3.2 does not take into account the directionality of the relationships. For example, if the relationship is "inhibits," then it is important to know which object is inhibiting the other object. The directionality finding process involves identification of the biological objects, the relationships, and the finally, the directionality. In this section, we describe each of these processes from a text mining context.

3.3.1 Identification of Biological Objects

We describe a hybrid method to address the specific challenges in object identification, where an object can be a gene, protein, cell type, organism, RNA, chemicals, disease, or drug. This consists of the following levels:

1. Use multiple dictionaries to identify known objects.
2. Use Hidden Markov Models (HMM) to identify unknown objects based on term suffixes, and,
3. Use N-Gram models to resolve object name ambiguity.

The tagging process begins with a Brill tagger (Brill, 1995) generating the POS. The process then continues with creating a dictionary of terms from databases (e.g. Swiss-Prot), for each class type (e.g. protein, gene, etc.) to be identified, and a dictionary such as WordNet (Brill, 1995), which contains the majority of other known nouns (e.g. lab, country, etc.) that may not be classified as a classified object. These dictionaries are to be single-token words, meaning they are very general in nature. To create these dictionaries, one would take a list of multi-token words (e.g. IkB inhibitor, RNA polymerase) which are defined in a class (e.g. protein) and then take the last word from each (e.g. inhibitor, polymerase). This can be described as $w_1w_2w_3...w_n \in$ MTD then $w_n \in$ STD where MTD is the multi-token dictionary, and STD is the single token dictionary, and w_i is the i^{th} term in a multi-token word. The other piece of data the user needs is a set of training documents from the area for which the objects are to be tagged.

Once the training data is obtained, two important steps are involved in the tagging process. First, an N-gram model describes a class (e.g. protein, gene, etc.) using the phrases of the surrounding context. Second, an HMM model describes a class based on the internal context. If abbreviations are present, then a separate HMM model is created to describe them using a separate dictionary for each class where abbreviations commonly occur. Protein and gene classes would need this extra model to fully describe them as they often use abbreviations.

N-Gram Models:

An N-Gram model is used to disambiguate object tags. An N-Gram model is described as a simple Markov model where the probability of a word W_l in position *n* can be given by the following equation (Jurafsky and Martin, 2000):

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_{k-N+1}^{k-1}) \tag{5}$$

where $P(w_k \mid w_{k-N+1}^{k-1})$ is the probability that w_k follows the previous *N* words. This is a simplification and assumes a word's probability is only dependent on the previous N characters. In order to calculate $P(w_k \mid w_{k-N+1}^{k-1})$ for each word in a given training corpus, the following general equation is used (Jurafsky and Martin, 2000):

$$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})} \tag{6}$$

where $C(w_{n-N+1}^{n-1} w_n)$ is the number of times the previous *N* words are followed by w_n and $C(w_{n-N+1}^{n-1})$ is the total number of times the previous *N* words occur. Often times a given corpus is not sufficient to encompass all words that may be encountered in a given corpus. It is necessary to use smoothing to help describe more accurately the probability of a given word. One of the best methods used is the Good-Turing method and is described as (Jurafsky and Martin, 2000):

$$c^* = \frac{(c+1)\dfrac{N_{c+1}}{N_c} - c\dfrac{(k+1)N_{k+1}}{N_1}}{1 - \dfrac{(k+1)N_{k+1}}{N_1}} \tag{7}$$

for 1 ≤ c ≤ k, where *c* is the original count of the word and N_C is the number of words counted *c* times.

Object Disambiguation Using N-Gram Models:

The N in N-Gram is the number of words in a given *pregram* or *postgram.* For this tagging method, the N-gram model takes the phrase data that was obtained using an N-Gram training process and uses it to define the probabilities for each class given a phrase. This probability is defined as:

$$P(c \mid phrase) = \frac{C(phrase_c)}{C(allphrases_c) + \sum_{i=0}^{M} C(phrase_i)} \qquad (8)$$

where *C(phrase$_c$)* is the number of *times* the phrase appears in class c, *C(allphrasesc)is* the number of times all phrases appear in class c, *M* is the total number of classes, and *C(phrase$_j$)* is the number of times the phrase appears in class *i.*

For N-Grams, that have N > 1, it becomes necessary to set up the model so that if the match for the full length N-Gram is not found, then the (N-l)-Gram can be tried, and if that does not work, the (N-2)-Gram, etc., would be needed. This stepping down can continue all the way down to the 1-Gram. If there is no match for the 1-Gram then the N-Gram fails to classify the object. The stepping down also allows the best possible N-Gram that matches to be found. Of course, a probability obtained using an N-Gram will always be greater than that obtained using an M-Gram where M < N. The following two-step process finds a class for a word having a *pregram*, *postgram*, or both:

1. Given a postgram and pregram,
2. Return the class c having the maximum $P(c_i \mid postgram) + P(c_i \mid pregram)$

If the probability for the class is zero, it shows that the word is not represented by the model and that it would require additional processing. Typically, smoothing would be done (e.g. Add-One Smoothing, Good-Turing Discounting, etc.; see (Jurafsky and Martin, 2000)).

Tagging of Abbreviations Using HMM:

An HMM is used to classify words that are abbreviations composed of less than six characters. The size of six was chosen based on the observation that most abbreviations are less than six characters long. This is done using a separate set of dictionaries specified by the user that are example abbreviations of several words known to fall into a specific class (e.g. SPF-l is a known protein abbreviation) that is known to contain abbreviations. These abbreviations are used as training data for the HMM. This abbreviated HMM will be referred to as the short HMM (SHMM). In addition, there may be longer words which are comprised of unusual symbols but represent an important object (e.g. Trpl53->Gly, which represents a specific change in a protein sequence). These longer words would have a separate HMM which will be referred to as the long HMM (LHMM).

An HMM contains states that represent a defined character type and the events in the states represent the specific characters in the word. To separate different words based on their characters, the model uses two groups of states. These states are identical in that they represent the same character types, but are different in that one represents a character type for a particular word type while the other represents the same character type for another word type. Figure 16-3 shows an example of different states of an HMM model to distinguish between words and gene names.

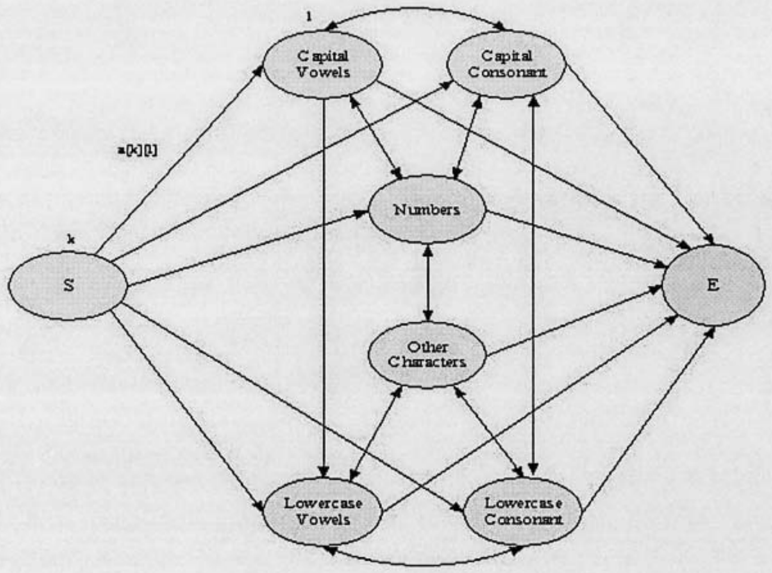


Figure 16-3. States of word tagger using HMM.

In this example, the state S represents the starting state. Within each state is the event probability of a given character occurring and is defined as $e[l](x[i])$ where *l* is the state and *x[i]* is the i^th^ character in sequence *x*. There is also a transition probability between each state showing the probability of going from one state to another defined as $a[k][l]$ where *k* is the state that is being left and *l* is the new state being visited.

The path is the sequence of states that occur and the probability of a given path for a sequence of characters is given by the sequence of states and the corresponding characters occurring in the sequence of characters. Formally, it can be expressed as follows (Durbin et al., 1998):

$$p(x,p) = a[0][1]\prod_{i=1}^{L} e[i](x[i])a[i][i=1] \tag{9}$$

The ending state of the most probable path is used to determine what object type the word is. In order to get the most probable path, the Viterbi algorithm (Durbin et al., 1998) is used.

3.3.2 Grouping Object Synonyms

The second stage in the whole process is resolving object synonyms correctly. The grouping of synonyms becomes complicated when (i) Synonyms share words and (ii) when Synonyms do not share words. Consider, for example, when they share words:

1. For the first time, **somatolactin (SL) cells** have ...
2. The **SL cells** were ...
3. The SL-immunoreactivity was mostly located in the granules of the **cells** ...

All three highlighted words in the above sentences refer to the same biological object. Knowing that the word *cells* in sentence three means the same thing as *SL cells* in sentence two and *somatolactin (SL) cells* in sentence one would have led to additional information that would have been specific enough for a biologist to use. Not knowing this would have caused the information extracted to be too general in the sense that the word "cell" by itself can represent more than one cell (e.g. somatolactin cell, heart cell, gonadotrope, etc.). The second case is when they do not share words:

1. Thyroid hormone receptors (T3Rs) are ...
2. T3Rs are bound by ...
3. **It** is found on ...
4.

Here the highlighted word "it" has little in common with the other two words. How to identify pronouns becomes important as information can be lost if they remain ambiguous.

In our approach, the word abbreviations are first processed through the PNAD-CSS algorithm (Yoshida et al., 2000). The abbreviations thus identified are used to group words together by merging words associated with the abbreviation with words associated with the full word from which the abbreviation was derived. In the first step in the grouping process, words are separated into their different classes (e.g. protein, gene). The next step is to build a generalized ontology and grouping of related words using a graph structure. The algorithms for this process can be broken down into two parts: the insertion of words from a group into a tree, and the extraction of word and their synonyms. The extraction process will create two categories of relationship between terms: one is a direct relationship like that of a word and its abbreviation, and the other is a hierarchical type where one word refers to several different words but those words do not refer to each other. For example, if someone talks about *proteins,* this is referring to more than one protein and not necessarily related, while if someone is referring to a *protein,* it is encompassing only one protein and a method similar to the pronoun tagging should be used.

A Grouping Example:

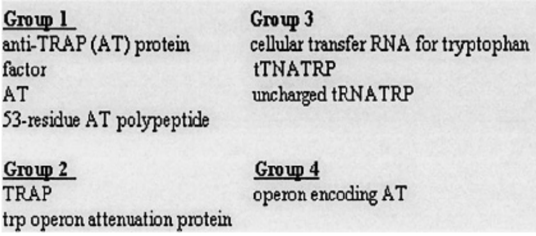
Consider the following text passage:

An anti-TRAP (AT) protein, a factor of previously unknown function, conveys the metabolic signal that the cellular transfer RNA for tryptophan (tTNA^{TRP}) is predominantly uncharged. Expression of the operon encoding AT is induced by uncharged tRNATRP. AT associates with TRAP, the trp operon attenuation protein, and inhibits its binding to its target RNA sequences. This relieves TRAP-mediated transcription termination and translation inhibition, increasing the rate of tryptophan biosynthesis. AT binds to TRAP primarily when it is in the tryptophan-activated state. The 53-residue AT polypeptide is homologous to the zinc-binding domain of DnaJ. The mechanisms regulating tryptophan biosynthesis in Bacillus subtilis differ from those used by Escherichia coli.

The tagging process would yield:

An **<p>anti-TRAP (AT) protein</p>**, a **<p>factor</p>** of previously unknown function, conveys the metabolic signal that the **<rna>cellular transfer RNA for tryptophan</rna>** (**<rna>tTNA^{TRP} </rna>**) is predominantly uncharged. Expression of the **<dna>operon encoding AT</dna>** is induced by **<rna>uncharged tRNA^{TRP}</rna>. <p>AT</p>** associates with **<p>TRAP</p>**, the **<p>trp operon attenuation protein</p>**, and inhibits its binding to its **<rna>target RNA sequences</rna>.** This relieves **<s>TRAP-mediated transcription termination</s>** and **<s>translation inhibition</s>,** increasing the rate of **<s>tryptophan biosythesis</s>. <p>AT</p>** binds to **<p>TRAP</p>** primarily when it is in the tryptophan-activated state. The **<p>53-residue AT polypeptide</p>** is homologous to the **<d>zinc-binding domain</d>** of **<p>DnaJ</p>.** The mechanisms regulating **<s>tryptophan biosynthesis</s>** in **<o>Bacillus subtilis</o>** differ from those used by **<o>Escherichia coli</o>.**

The Grouping Process will then generate:



Group 1
anti-TRAP (AT) protein
factor
AT
53-residue AT polypeptide

Group 2
TRAP
trp operon attenuation protein

Group 3
cellular transfer RNA for tryptophan
tTNATRP
uncharged tRNATRP

Group 4
operon encoding AT

It can be observed that grouping these words can greatly change the statistical nature of the terms when one word in the group is used for all other words in the same group, helping methods to achieve accurate statistical measures.

3.3.3 Extracting Object Relationships

The final step in the process is to extract relationships between the tagged entities. This process defines two types of relationships. The first is referred to as *directional relationships.* These relationships include for example, *protein A inhibits protein B.* In this case, a biologist not only needs to know what the relationship is but also the direction in which the relationship occurs. The second type of relationship is referred to as hierarchical relationships. This type of relationship would include, for example, *the brain is part of the nervous system.* The next two sections discuss the techniques used for each type of relationship.

Directional Relationships:

Directional relationships are found using a Hidden Markov Model. This is accomplished by generalizing words based on their POS tag or object tag so that the model encompasses a wide variety of relationships without having a lot of training data. The idea behind the scheme is that each relationship has a certain form. When a sentence is given to the model, its state sequence will contain the states created for a specific relationship classifying the sentence to be that relationship. The direction is detected by creating two event sequences, one where one of the objects in the relationship is classified as the subject while the other sentence has the other object classified as the subject. The model would then give the highest probability to the sentence with the correct subject, indicating the direction of the relationship. This is important as there may be instances where in one sentence the subject comes before the verb (e.g. protein A binds protein B) and in another it comes after the verb (e.g. protein B is bound by protein A).

To understand this model, a short example showing how the model is built for two sentences representing the same relationship but having a different position for the subject and object is shown. The two sentences are as follows:

1. Protein A inhibits Protein B.
2. Protein B is inhibited by Protein A.

The sentences are then Brill tagged and the objects identified as:

1. *<protein>Protein A</protein> inhibits/VBZ <protein>Protein B</protein>*
2. *<protein>Protein B</protein> is/VBZ inhibited/VMX by/IN <protein>Protein A</protein>*

All possible relationships (e.g. protein-protein) that were defined by the user (during training) as directional are then extracted from the sentence to give the following possibilities:

1. Possibilities for sentence 1: <subject>Protein A</subject> <object>Protein B</object>, <object>Protein A</object> <subject>Protein B</subject>
2. Possibilities for sentence 2: <subject>Protein B</subject> <object>Protein A</object>, <object>Protein B</object> <subject>Protein A</subject>

A user would then define the type of relationship each sentence is and the relationship with the correct labeling. The produced event and state sequences needed to train the HMM would be as follows:

1. Class for sentence 1: Inhibits
   Event sequence: [subject][inhibits][object]
   State sequence: [subject][/VBZ][object]
2. Class for sentence 2: Inhibits
   Event sequence: [object][is][inhibited][by][subject]
   State sequence:[object][/VBZ][/VMX][by/IN][subject]

Once the event and state sequences are known, the parameters of the HMM are determined using the following equations (Durbin et al., 1998):

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \qquad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \tag{10}$$

where a_{kl} is the probability of a transition from state k to state l, $e_k(b)$ is the probability that an event b occurs in state k, A_{kl} is the number of transitions from state k to state l in the training data, and $E_k(b)$ is the number of times the event b occurs in state k in the training data. Once the model is trained, the relationship finding using the model can be carried out. For example, consider the following sample sentence:

1. ADH is inhibited by alcohol.

The sentence is tagged and the possible relationships extracted. Assuming one of the defined directional relationships is between a chemical and protein, the process produces the following two event sequences:

1. [subject] [is] [inhibited] [by] [object]
2. [object] [is] [inhibited][by][subject]

The model generates the probability of event sequence 1 as 0.0 while the probability of event sequence 2 as 0.0625 (the actual algorithm is omitted here for the sake of brevity). Taking the higher probability, this produces the relationship that *alcohol* inhibits *ADH* as opposed to *ADH* inhibits *alcohol.* The general process for producing an event sequence for any sentence which is tagged can be given by the following process.

1. If the word is a tagged object, make the event the object tag.
2. If the word is not an object, make the event the word.
3. For each possible directional relationship found, produce two event sequences where each object in the relationship is represented as the subject event, each object is also represented as an object event, and each sequence has one subject event and one object event.

One advantage of this method is that it avoids the complex issues of creating rules encompassing all possible relationships that are needed in a rule-based approach. In addition, generalizing objects allow for more flexibility in the model and enables the detection of new relationships without having to define a specific event probability for an object which is not specified by the model in its unclassified form. Another advantage for using the HMM for classification includes the ability to overcome noise by allowing default event probabilities in cases where an event may not be defined. This allows sentences to be classified to their most probable classification despite the HMM not having seen the event sequence previously in training. In addition, the verb states (e.g. /VBZ, /VMX) can be modified to include new verbs which define new directional relationships.

Hierarchical Relationships:

The Hierarchical relationships take advantage of the fact that verbs are not important to the relationship. Hence, this type of relationship can be defined in purely statistical terms using only the parent and child of the relationship. This technique is closely related to the co-occurrences of the gene extraction process described in Section 3. This is different because before the association matrix was square and considered relationships between two objects that were classified in the same class. Now the relationship is defined in such a way that it considers relationships between any objects regardless of what class they are classified in.

3.3.4 Experimental Results

Various experiments were carried out to evaluate the performance of the system at all three stages of the process, namely, tagging, grouping and relationship extraction. Results on each are considered separately below.

Tagging Performance:

The tagging method was applied to 100 abstracts from Medline obtained using the keyword "pituitary." The results were quantified using the measurements precision, recall, and F-Score defined earlier. The N-Gram model's default phrase length was three, making it a 3-Gram model. The training data used for the 3-Gram was comprised of 2,000 abstracts obtained from Medline using the keyword phrase "protein interaction."

The tagging performance using the dictionary only was only 50-60% despite using a large dictionary of words extracted from Swiss-Prot. The addition of the HMM and N-gram to the tagging process produced the results in Table 16-4 and has an average F-Score of 70%. The final step of the tagging process which made corrections for mis-tagged abbreviations greatly increased specificity and recall by eliminating false positives from the HMM tagged protein and gene objects, and increased recall for other object types as their formally mis-tagged abbreviations are tagged correctly.

Experiments were also conducted by increasing the length of the 3-Gram model to a 4-Gram model, and only the protein object tagging performance increased slightly in both precision and recall. This can be expected considering the low number of 3-Grams found in the 3-Gram model, which would indicate an even lower number of 4-grams. Add to this the fact that the majority of 3-Grams that were found were surrounding the protein object type, it would be expected that the new 4-Grams found would most likely effect protein object tagging.

Table 16-4. Results of Gene/Protein Classification

Tag Type	Correct	Missed	Recall	Precision	F-Score
Protein	533	150	78%	67%	72%
Gene	54	66	45%	57%	50%
Chemical	115	44	72%	69%	71%
Organism	305	130	70%	76%	72%
Organ	171	96	64%	81%	72%
Disease	202	93	68%	60%	64%

Performance of Grouping:

Performance of the grouping process used the same set of documents used for evaluating the tagging performance. A good way to measure the grouping performance is to see how it reduces the amount of information in terms of unique objects. The grouping showed a drop in the number of protein-protein relationships, cell-protein relationships, and organ-cell relationships. The grouping of synonyms thus greatly reduced the complexity of the data and helped objects to become more specific. The number of unique objects dropping by 24.7% indicates that an object may be written in many different ways. This is particularly true when looking at protein names. The drop of 38% in the relationships between proteins, due to grouping, directly shows the way in which the same protein takes on many different word forms. This drop is less when referring to the relationships between proteins and cells, indicating that the use of different names to refer the same cell object are much less common than that of proteins. This is further illustrated where the drop is only 2% for binary relationships between organs and cells. This would indicate that the use of different names for organs is almost non-existent. These results are expected as proteins are much more likely to take on different names in a document than a cell type or organ.

The overall performance of grouping was obtained by going through ten grouped abstracts and counting the number of terms grouped correctly and grouped incorrectly. These results are shown in Table 16-5.

Table 16-5. Performance of Grouping process

# of Correct group terms	# of Missed group terms	# of Incorrect group terms	Recall	Specificity
46	4	10	92%	82%

Results on Object-Object Relationship Extraction:

Directional Relationships: The HMM model was first trained using four directional relationships: inhibit, activate, binds, and same for the problem of protein-protein interactions and were trained for the directional HMM. When the model was used on the training set of sentences, it was found to have recorded the relationships with a recall and specificity of 100%, which would be expected given the small number of relationships.

The model was then tested against a larger corpus of text consisting of 1,000 abstracts downloaded from Medline using the key word of "protein interaction." These abstracts were then tagged and grouped, and all possible protein-protein interactions were extracted without specification to direction. The possible relationships were then passed through the trained HMM to extract directional relationships. In all, there were 53 such relationships extracted of which 43 were correct giving a specificity of 81%.

Hierarchical Relationships: To test the hierarchical relationships, the same pituitary corpus that was used to test grouping was used. Of these, 83 were specific and accurate enough to be useful while 49 were either wrong or too general to be useful giving a specificity of 65%. Having a threshold on the association value of 20 would change it to be 57 and 14, respectively, giving a specificity of 82%. More details about this work can be found in (Palakal et al., 2002c; Palakal et al., 2003).

4. BIOMAP: A KNOWLEDGE BASE OF BIOLOGICAL LITERATURE

In this section we present the progress made to develop a complete "knowledge base" of associations between biological objects that are important for biologists to study and understand specific biological processes. The main goals of this effort are (a) to develop a very large knowledge base, called BioMap, using the entire Medline collection of literature documents (over 12 million), and (b) to develop an interactive knowledge network for users to access this secondary knowledge (BioMap) along with its primary databases such as Medline, GenBank, etc., in an integrated manner based on a specific area of problem enquiry. The development of BioMap and its associated access "window" (the knowledge network), all of the text mining tasks that were discussed in the previous sections (such as identification of biological object names and discovering object-object relationships) will be utilized. The overall architecture of BioMap is shown in Figure 16-4.

The BioMap system basically consists of a set of organism-specific Knowledge Bases, a collection of intelligent algorithms for biological Object Tagging, Identification, and Relationship Discovery, System Interface, and a User Interface. A multi-level hybrid approach that incorporates statistical, stochastic, neural network and N-Gram models along with multiple dictionaries are used to handle the multi-object identification and relationship extraction problem for BioMap as described in Section 3.

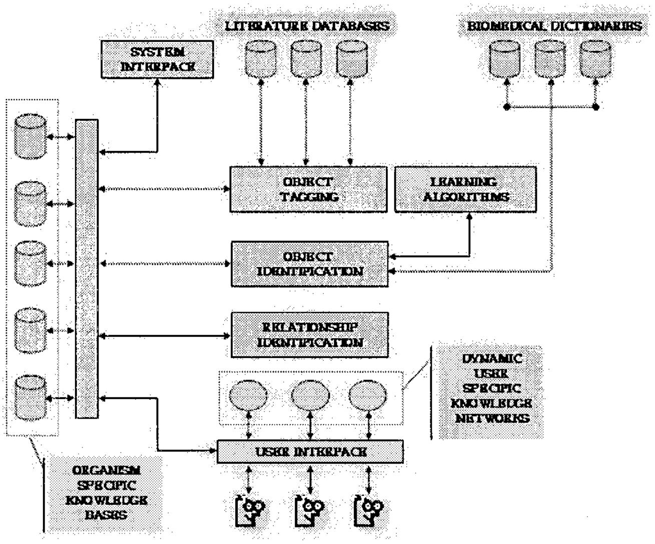


Figure 16-4. The overall organization of the BioMap System

The relationships thus discovered from the Medline collection are maintained in a relational database along with the specific links to literature sources, genes and protein sequence databases, popular bioinformatics tools such as PubGene, GO, etc., as well as links to image sources if the proof of relationships appear as images in the literature (as in the case of microarray experimental results). A user can access this knowledge base using any simple or complex queries, a disease name, a set of gene names, or any such combinations. Unlike in traditional databases, the outcome of a user query will be a complex set of data with multiple associations among them. Furthermore this network will be viewed in a hierarchical manner, allowing biologists to transcend the molecular view and see the physiological context from which a relationship is pulled. An example of the constructed knowledge network (knowledge view of BioMap) is shown in Figure 16-5. Knowledge outside of the hierarchical view can be pulled in, but the biologist will be able to specify what the context of the knowledge brought in is. For example, a biologist may start out looking at protein interactions shown to occur in the hippocampus of the human brain. The user may then choose to bring in additional interactions from the hippocampus of the rat brain. Understanding the context in which different interactions are playing a role is a key in understanding the function of a biological object. Different context often means the biological object can have a different function as described recently in (Brill, 1995). This context view is often overlooked.

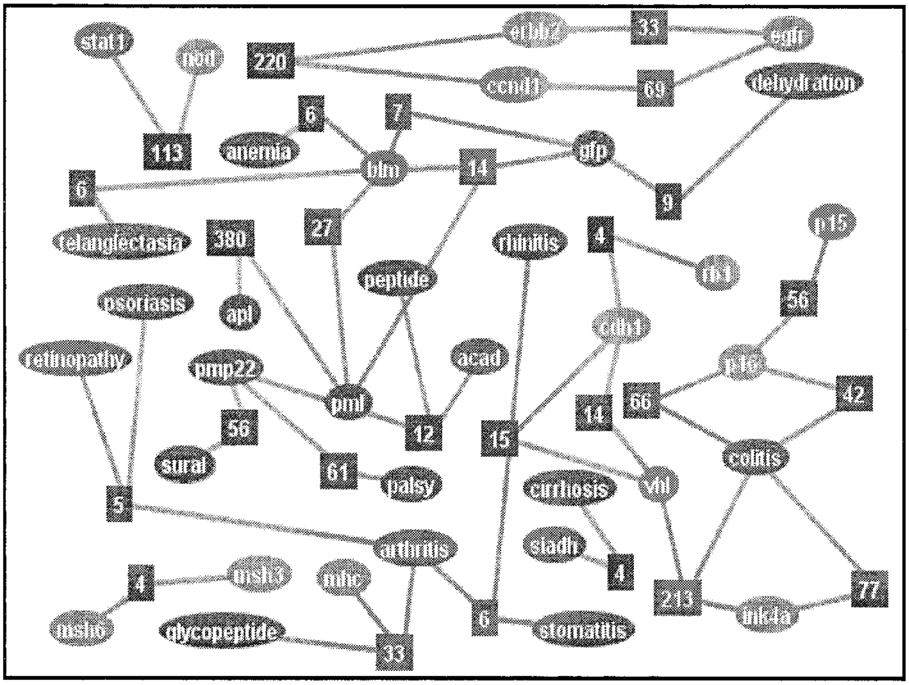


Figure 16-5. BioMap Knowledge represented as Hypergraph

The knowledge network is constructed as a hypergraph "on the fly" based on each user query. A hypergraph is an extension of a graph in the sense that each hyperedge can connect more than two vertices, thus allowing connecting relationships among multiple objects simultaneously. A system based on the hypergraph model has a number of advantages: the model is independent from updates to the underlying database; it enables the formation of hypergraphs from entities in different databases; it allows the system to be accessed by multiple users simultaneously, each with an independent hypergraph; further queries can be made to a hypergraph to obtain a better "focused" view of the knowledge base; it reduces the need to access the knowledge base multiple times when a hypergraph is to be shown to a remote user; and most importantly, the interaction with the user can be made faster when a query is made on the hypergraph since it is available in the local memory. Furthermore, the edges and vertices of the hypergraph will be "live" (made as hyperlinks) allowing the user to access primary data sources and bioinformatics tools pertinent to the information contained in the knowledge network.

In order to develop such a derived knowledge-base of associations from the primary source of biological literature databases, several research issues need to be resolved and these are described in the chapter. These include object identification (tagging), ambiguity resolution, synonym resolution, abbreviation resolution, and finally associations discovery and visualization. The knowledge network is a "window" to the BioMap's large knowledge base. Unlike traditional *binary relationships* among objects, BioMap's knowledge base has rich *multi-way relationships* such as that captured by the sentence "gene A inhibits protein B in pathway C in the context of disease D in organ E." This naturally leads to ternary, quaternary or even higher-order relationships and hence, to the notion of a hypergraph. A hypergraph is a generalization of a binary relationship graph (as described in the chapter) and is characterized by G = (V,E) where V is the set of vertices and E is the set of hyperedges. Unlike regular graphs where elements of E are pairs of vertices, denoting binary relationships, a hyperedge in a hypergraph is a subset of V and corresponds to a multi-way relationship of (possibly) more than two objects included in the subset. As in the case of binary edges, the multi-way associations (hyperedges) can be determined by co-occurrence based mining from BioMap's knowledge base. It is clear that the number of such possible hyperedges is combinatorially exponential with the number of objects, since, the number of subsets of a set A of cardinality n (i.e., the cardinality of the power set of A) is 2^n. This is in contrast to the binary graph, encoding binary relationship, where the number of possible associations (edges) is quadratic in n. Hence, any exhaustive attempt to check for all hyperedges will run into extremely high computational complexity, particularly since the total number of objects in the entire BioMap knowledge base is expected to be very large (in the hundreds of thousands). Hence, heuristic approximations are needed to limit the number of possible hyperedges.

4.1 BioMap Knowledgebase

The BioMap knowledgebase can be viewed as one large database or it can be conceptually divided by a concept such as organism to allow for greater scalability. The schema for the database is shown in the Figure 16-6. The Noun_Phrases table stores the extracted noun-phrases from text. This table is used for classifying these noun-phrases into different biological objects using various sources like UMLS, LocusLink etc. and machine learning techniques. These classified noun phrases are identified in the table Classified_Noun_Phrases that contains the type of object represented (e.g. organism) by this noun-phrase and the method used to classify it (e.g. dictionary), which are in turn stored in Categories and Methods tables. Each noun phrase can then be associated with a defined object through the Defined_Noun table. Relationships between defined objects can then be stored in the Relationship table. Within the relationship table there is a typical binary relation between two objects but also dependencies on that relationship can be described using the Relationship and Object dependency tables. The dependency can be either positive or negative. In this way, more complex relationships can be described beyond the binary relationships. The Complex_Relationships table adds to this by creating objects made up of relationships between other objects (e.g. a protein complex is made up of binding relationships between proteins and in some cases RNA)

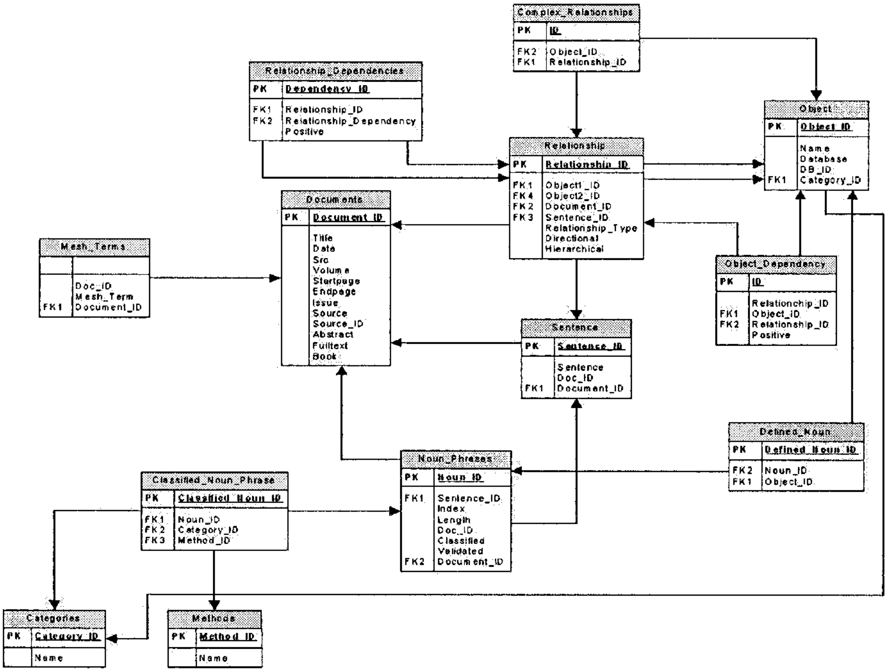


Figure 16-6. BioMap database schema

The BioMap knowledge base is implemented using Oracle 9i databases. The documents for each database are acquired from Medline. Once the database is populated and the noun-phrases are classified using different methods, they will form a basis for the knowledge network as discussed in the previous section, and they provide readily available data for testing and employing new techniques for object name resolution and other interesting text mining problems.

A major step in the creation of the knowledge base is to populate the database with relevant information from the text documents. This process involves identifying objects such as Gene, Protein, Cell Type, Organ, Organelle, Chemical/Drug, and Disease, and was carried out using the multi-level approach described in the previous section.

In our prototype study, we used a set of 30,000 documents and used UMLS and LocusLink to identify and classify objects. UMLS is a major source to resolve the noun-phrases into a number of categories. As a second step to improve on direct matches to UMLS concepts, the MetaMap Transfer (MMTx) program was used to map the noun-phrases to UMLS concepts. The MetaMap Transfer API in Java is used for this purpose. Again, only unambiguous matches are considered that also give a maximum score of 1,000. A score of 1,000 means that the mapping MetaMap came up with is the best one. The default parameters for the MetaMap are used. Overall, UMLS classifies the entities into a number of categories, which include genes, proteins, drugs/chemicals, and diseases, among others. The reason for the two steps used for UMLS is the following. Doing a direct match is much faster and the majority of entities resolved by the two steps is covered by the first step of direct comparison. MetaMap is used as an important second step to catch those noun-phrases that are similar but are not exact matches to a UMLS concept. LocusLink is then used for classifying gene names. LocusLink is a resource provided by NCBI that provides "genecentric" information for various organisms. LocusLink is particularly suited to the task as it has genetic information for multiple organisms. Currently "human," "rat" and "mouse" are being used to create BioMap. For each database for human, rat, and mouse, respective dictionaries of gene names are created from LocusLink. These dictionaries are then used to resolve the gene names in each respective organism's database. The noun-phrases that have not been resolved by UMLS are looked up in the LocusLink gene dictionary. If a match is found, then that entity is classified as "gene."

4.2 Results and Discussions

The results for entity name resolution for documents relating to human, rat and mouse are presented here. These results are based on the databases created using 30,000 documents from Medline and resolving noun-phrases using UMLS and LocusLink. The results are summarized in Table 16-6.

Table 16-6. Results of Name Resolution using Dictionaries

Noun-phrases		UMLS	LocusLink	Total
Human	Total	789,551		
	Classified	217312	9561	226873
	Percentage	27.52%	1.21%	28.73%
Rat	Total	94,212		
	Classified	21408	1261	22669
	Percentage	22.72%	1.34%	24.06%
Mouse	Total	89422		
	Classified	21385	2139	23524
	Percentage	23.91%	2.39%	26.31%

UMLS is a major contributor for resolving the object names followed by LocusLink. UMLS has resolved the entities into 132 distinct categories. LocusLink has resolved roughly 2% of the total nounphrases into genes.

As an example, from the experiments we have done, we consider some of the noun-phrases and walk through the process. In the first step of classification using UMLS, the noun "apoptosis" is classified as a "Cell Function." After that in the next step when we applied the MetaMap Transfer method, the noun-phrase "urinary infection," which was missed by the direct match method was mapped to "Urinary infection NOS (Urinary tract infection)" concept in UMLS, which belongs to a "Disease or Syndrome" category in the UMLS Metathesaurus. In the last step using LocusLink, let us consider the noun "FADD." The noun "FADD" was not classified by either UMLS methods, but LocusLink classified it as a gene.

As discussed in the previous section and evident from sample results, a multilevel approach to resolving names is quite effective in identifying important entities using a specialized dictionary for those types of entities. However, as we can see from the above results, the dictionaries can resolve only up to 30% of the nouns. This may probably be improved to 40% by using more specialized dictionaries for more types of entities. However, there is only so much that can be achieved using only a dictionary lookup approach. There is clearly a need for sophisticated algorithms to successfully classify the entities. The machine-learning techniques such as Hidden Markov Models (HMM) and N-grams to tackle the entities left unresolved by the dictionary look up approach is currently being developed.

The key innovative features of the proposed BioMap are that it is adaptable and scalable, and that the core knowledge base will be constructed from a very large collection of documents (Medline) to make the system robust. The adaptation feature requires that the system should have the ability to learn new problem domains without having to rebuild the system as in the case of fully rule-based or grammar-based approaches. The scalability feature allows the system to continue to develop its knowledge base as new information arrives in the literature databases or incorporating information from other data sources (e.g. Science, Nature, etc.). BioMap is novel in its ability to transcend typical views of data that only consider a small scope of objects and relations and allows for a global view of interactions among objects. It is hoped that this view will help biologists transcend the bottlenecks that keep them from relating findings at the molecular level to real physiological changes which characterize disease. Further discussions on BioMap can be found in (Kumar et al., 2004).

5. CONCLUSIONS

The biological literature databases continue to grow rapidly with vital information that is important for conducting sound biomedical research. The objective of the research described in this chapter is to develop, for the first time, a scalable knowledge base (BioMap) of biological relationships from vast amount of literature data. The results of this research will significantly enhance the ability of biological researchers with diverse objectives to efficiently utilize biomedical literature data. BioMap will be a new type of "secondary" knowledge resource derived from primary resources such as Medline. It will be the "window" to every biomedical researcher who will be seeking knowledge from the literature databases, however, without being overwhelmed by its large volume.

When the knowledge network is presented to the user as a rich hypergraph, it enables easier browsing of the content shown to the user. Each node in the hypergraph can be made to be rich in content. They can be clicked on to show a new hypergraph dynamically with the selected entity as the seed. They contain information such as citations from where the corresponding term is derived. The biological objects such as genes, proteins, drugs, etc., will be represented as nodes in a hypergraph. Hence, BioMap will not only be an effective aid for biomedical research, but also a teaching and learning tool for high school, undergraduate, and graduate students pursuing academic programs in biomedical sciences. Another significant contribution of this work is the ability of the system to efficiently discover associations not explicitly reported in any one document, but based on associations implicitly hidden in multiple documents.

6. ACKNOWLEDGEMENTS

The research reported in this paper was supported in part by National Science Foundation Information Technology Research grant number NSF-IIS/ITR 0081944, NIH BISTI grant, NIH-NIGMS P20 GM66402, and by an NSF/NSDL grant NSF-DUE-0333623. The computing facilities used for this research were supported in part by the Indiana Genomics Initiative (INGEN). The Indiana Genomics Initiative (INGEN) of Indiana University is supported in part by Lilly Endowment Inc.

REFERENCES

Ashburner, M. (2000). "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, vol. 25, 25-29

Atschul, S.F., Gish, W., Miller, W., Myers, E., Lipman, D. (1990). "Basic Local Alignment Search Tool," *Journal of Molecular Biology* 215, 403-410

Baasiri, R.A., Glasser, S.R., Steffen, D.L., Wheeler, D.A. (1999). "The Breast Cancer Gene Database: A Collaborative Information Resource," *Oncogene* 18:7958-7965, http://tyrosine.biomedcomp.com/4d.acgi$tsrchname?Name=&topic=BCIR

Brill, E. (1995). "Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," *Computational Linguistics*, 21 (4):543-566

Chinchor, N. (1995). "MUC-5 Evaluation Metrices," in *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, Baltimore, Maryland, USA, 69- 78

Collier, N., Nobata, C., and Tsujii, J. (2000a). "Extracting the Names of Genes and Gene Products with a Hidden Markov Model," *Coling 2000*, 201-207

Craven, M., Kumlien, J. (1999). "Constructing Biological Knowledge Bases by Extracting Information from Text Sources," *ISMB* : 10 – 20

Durbin, R., Eddy, S., Krogh, A., Mitchison, G. (1998). *Biological Sequence Analysis.* Cambridge University Press. New York, N

Friedman, C., Kra, P., Yu, H., Krauthamrner, M., Rzhetsky, A. (2001). "Genies: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles," *Bioinformatics* 17 Suppl. 1, S74 -S82

Fukuda, K., Tsunoda. T., Tamura, A., and Takagi, T. (1998). "Toward Information Extraction: Identifying Protein Names from Biological Papers," in *Proceedings of the Pacific Symposium on Biocomputing*, 705-716

Genecards, http://mach1.nci.nih.gov/cards/index.html

Hatzivassiloglou, V., Duboue, P., Rzhetsky, A. (2001). "Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach," *Bioinformatics*, 17 Suppl. 1, S97 -S106

Honderich, T. (1995). *The Oxford Companion to Philosophy*, Oxford University Press, http://www.xrefer.com/entry/553381.

HUGO, http://www.gene.ucl.ac.uk/nomenclature/

Humphreys, K., Demetrios, G., and Gaizauskas, R. (2000). "Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures," in *Proceedings of the Pacific Symposium on Biocomputing*, 505-516

Iliopoulos, I., Enright, A.J., and Ouzounis, C.A. (2001). "Textquest: Document Clustering of Medline Abstracts for Concept Discovery in Molecular Biology."

Jenssen, T.K., Laegreid, A., Komorowaki, J., Hovig, E. (2001). "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression," *Nature Genetics*. May 28 (1); 21-8, http://www.pubgene.org/

Joshi, A.K. (1998). "Role of Constrained Computational Systems in Natural Language Processing," *AI Journal*, 103, 117-132

Jurafsky, D., and Martin, J. Speech and Lanuae (2000). *Processing*. Prentice-Hall, Inc. Upper Saddle River, New Jersey

Kazama, J., Makino, T., Otha, Y., Tsujii, J. (2002). "Tuning Support Vector Machines for Biomedical Named Entity Recognition," http://www.snowelm.com/~t/research/pub/./kazama_aclbio02.pdf

Krauthammer, M., Rzhetsky, A., Morozov, P., Friedman, C. (2000). "Using BLAST for Identifying Gene and Protein Names in Journal Articles," *Gene* 259: 245 – 252

Kumar, K., Palakal, M., and Mukhopadhyay, S. (2004). "BioMap: Toward the Development of a Knowledge Base of Biomedical Literature," in *2004 ACM Symposium on Applied Computing,* Nicosia, Cyprus

Leroy, G. and Chen, H. (2002). "Filling Preposition-Based Templates to Capture Information from Medical Abstracts," in *Proceedings of the Pacific Symposium on Biocomputing* 7, 350-361

LocusLink, http://www.ncbi.nlm.nih.gov/LocusLink/

Lodish, H., Berk, A., Matsudaira, P., Baltimore, D., Zipursky, S., Darnell, J. (1995). *Molecular Cell Biology.* Third Edition. Scientific Books, Inc. New York

Marcotte, E.M., Xenarios, I., and Eisenberg, D. (2001). "Mining Literature for Protein-Protein Interactions." *Bioinformatics*, 17: 359-363

MedMiner, http://discover.nci.nih.gov/textmining/filters.html

MetaMap Transfer, http://mmtx.nlm.nih.gov

Ng, S. and Wong, M. (1999). "Toward Routing Automatic Pathway Discovery from On-line Scientific Text Abstracts." *Genome Informatics*, 10:104-112

Nobata, C., Collier, N., and Tsujii, J. (2000). "Automatic Term Identification and Classification in Biology Texts," in *Proceedings of the Natural Language Pacific Rim Symposium (NLPRS '2000)*, 369- 375

OMIM, http://www3.ncbi.nlm.nih.gov/htbin-post/Omim/

Ono T., Hishigaki H., Tanigami A., and Takagi T. (2001). "Automatic Extraction of Information on Protein-Protein Interactions from the Biological Literature," *Bioinformatics*, 17(2):155-161

Oyama T., Kitano K., Satou K., and Ito T. (2002). "Extraction of Knowledge on Protein-Protein Interaction by Association Rule Ddiscovery," *Bioinformatics*, 18(5):705-714

Palakal, M., Stephens, M., Mukhopadhyay, S., Raje, R. and Rhodes, S. (2003). "Identification of Biological Relationships from text documents using efficient computational Methods," *Journal of Bioinformatics and Computational Biology*, Vol. 1(2), 1-34

Palakal, M., Stephens, M., Mukhopadhyay, S., Raje, R. (2002c). "A Multi-level Text Mining Method to Extract Biological Relationships," in *CSB2002*, Stanford, CA

Park, J.C., Kim, H.S., and Kim, J.J. (2001). "Bidirectional Incremental Paring for Automatic Pathway Identification with Combinatory Categorical Grammar," Oct., http://citeseer.nj.nec.com/384291.html

Proux, D., Rechenmann, F., and Julliard, L. (2000). "A Pragmatic Information Extraction Strategy for Gathering Data on Genetic Interactions," in *Proc Int Conf Intell Syst Mol Biol* 8, pages 279-285

PubGene, http://www.pubgene.com/

PubMed, http://www.ncbi.nlm.nih.gov/entrez

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D. (1997). "GeneCards: Encyclopedia for Genes, Proteins and Diseases," Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel), http://bioinformatics.weizmann.ac.il/cards

Rindflesch, T.C., Tanabe, L., Weinstein, J.N., and Hunter, L., (2000). "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature," in *Proceedings of the Pacific Symposium on Biocomputing*; 517-28

Rothblatt, J., Novick, P., Stevens, T. (1994). *Guidebook to the Secretory Pathway.* Oxford University Press Inc., New York

Salton, G. 1989, Automatic Text Processing. Addison-Wesley

Sanchez C., Lachaize C., Janody F., Bellon B., Roder L., Euzenat J., Rechenmann F., and Jacq B. (1999). "Grasping at Molecular Interactions and Genetic Networks in Drosophila Melanogaster Using Flynets, an Internet Database," *Nucleic Acids Res,* 27(1):89-94

Smalheiser, N.R. (2002). "Informatics and Hypothesis-driven Research," *EMBO Rep* 3:702

Smalheiser, N.R. (2001). "Predicting Emerging Technologies with the Aid of Text-Based Data Mining: A Micro Approach," *Technovation* 21: 689-693

Steffen, B., Ulf, B., Faulstich, L., Hakenberg, J., Leser, U., Plake, C., Scheffer, T. (2003), http://www.pdg.cnb.uam.es/BioLink/workshop_BioCreative_04/handout/pdf/user12_1a.pdf

Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R., Mostafa, J. (2001). "Detecting Gene Relations from Medline Abstracts," in *PSB* 2001: 483-495

Swanson, D.R. and Smalheiser, N.R. (1997). "An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery," *Artificial Intelligence* 91: 183-203

Tan, A-H. (1999). "Text Mining: The State of the Art and the Challenges," in *Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*, 65-70

Tanabe, L. and Wilbur, W.J. (2002). "Tagging Gene and Protein Names in Biomedical Text," *Bioinformatics*, 18(8): 1124-1132

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll M. (2000). "Automatic Extraction of Protein Interactions from Scientific Abstracts," in *Proceedings of the Pacific Symposium on Biocomputing*, 541-551

UMLS, http://www.nlm.nih.gov/research/umls/umlsmain.html

Warshall, S. (1962). "A Theorem on Boolean Matrices," *JACM* 9:11-12,

Yakushiji, A., Tateisi, Y., Tsujii, J., Miyao, Y. (2000). "Use of a Full Parser for Information Extraction in Molecular Biology Domain," *Genome Informatics* II: 446-447

Yoshida, M., Fukuda, K., and Takagi, T. (2000). "PNAD-CSS: A Workbench for Constructing a Protein Name Abbreviation Dictionary," *Bioinformatics,* 16, 169-175

SUGGESTED READINGS

Hearst, M.A. (1998). *Automated discovery of wordnet relations.* In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA
This chapter discusses methods to extract general lexico-semantic relationships (e.g., x is a kind of y) by extracting corresponding patterns from text. Some biological relationships could also be of lexico-semantic nature.

Salton, G. (1983). *Introduction to Modern Information Retreival.* McGraw-Hill, New York
A popular authentic textbook on information retrieval techniques, including discussion of vector-space (tf-idf) and other models of text retrieval.

Ashburner, M. (2001). Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. *Genome Research*, 11, pp. 1425-1433
The gene ontology consortium provides a database of a large number of gene names and their relationships to biological functions, processes, and cellular locations. While this exercise is useful in its own right, this also can be used for tagging and pathway generation in biological context.

Hirschman L, Park JC, Tsujii J, Wong L, Wu CH. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics.* 18(12):1553-1556
This paper provides a review of text mining in biology. While reviewing several works in object tagging, relationship extraction, and pathway generation, it discusses the requirements of benchmark information extraction task datasets for biological text mining.

Scheffer, T. (2004). *Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, Pisa, Italy, September, Available on-line at: http://www.informatik.hu-berlin.de/Forschung_Lehre/wm/ws04/
A collection of papers describing some very recent research in text and data mining for Bioinformatics. There are papers on extracting protein-protein interactions and protein-function relationships, as well as other applications of text mining.

ONLINE RESOURCES

GENECARD: Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1997).
GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel).
http://bioinformatics.weizmann.ac.il/cards

Medline
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

PIR (Protein Information Resources)
http://pir.georgetown.edu/

SUN's GRAPH VIEWING APPLET:19. Sun Microsystems, Inc. (1995)
Graph.java demonstration software. Sun Microsystems Inc.
http://java.sun.com/applets/jdk/1.0/demo/GraphLayout/index.html

TRANSMINER
http://sifter.cs.iupui.edu/~sifter/transMiner/TransMinerBCResults.html

QUESTIONS FOR DISCUSSION

1. One of the problems in association discovery is determining the direction, if any, of a particular association. What can be some of the approaches in determining such directionality through text mining?

2. Some of the approaches that can be used for object identification in text include rule (grammar) based, statistical, and connectionist or other machine learning approaches. What are the relative advantages and disadvantages of the different approaches?

3. Since there seems to be the possibility of a variety or a bank of multiple taggers (object identifiers) designed using possibly different computational techniques, a question arises as to whether it is possible to improve the tagging performance further by combining them in a judicious way. What are some of the issues involved in designing such a meta-tagger?

4. The general hypergraph construction algorithm is believed to be computationally very complex. Why? What could be some of heuristics and/or approximations that can be used to make it more tractable?

5. Information visualization: The user-specific knowledge graph (or, hypergraph) can be quite complex involving a large number of nodes and associations. What could be some approaches to visualizing such large graphs in a cognition-rich manner?