**Hagit Shatkay**
is an associate professor at the School of Computing, Queen's University. She works in the area of machine learning and its application to biomedical data. She is an active member of the biomedical text-mining community, and one of the first researchers in the area of text mining and information retrieval for bioinformatics.

# Hairpins in bookstacks: Information retrieval from biomedical text

*Hagit Shatkay*

Date received (in revised form): 5th July 2005

## Abstract

Current advances in high-throughput biology are accompanied by a tremendous increase in the number of related publications. Much biomedical information is reported in the vast amount of literature. The ability to rapidly and effectively survey the literature is necessary for both the design and the interpretation of large-scale experiments, and for curation of structured biomedical knowledge in public databases. Given the millions of published documents, the field of information retrieval, which is concerned with the automatic identification of relevant documents from large text collections, has much to offer. This paper introduces the basics of information retrieval, discusses its applications in biomedicine, and presents traditional and non-traditional ways in which it can be used.

***Keywords:*** *information retrieval, text mining, biomedical text mining, biomedical literature mining, information retrieval, NLP*

## INTRODUCTION

The past decade has been marked by an unprecedented growth in both the production of biomedical data and the amount of published literature discussing it. The ultimate goal of modern large-scale biology is to translate these large amounts of data into actual knowledge of the complex biological processes and accurate models of living cells and organisms.

Any known or postulated information pertaining to entities such as genes, proteins, disease, drugs and their role in biological processes is published in the literature. The advancement of genomic and proteomic technologies is accompanied by an overwhelming increase in the number of publications discussing genes and proteins. The abundance of both biological data and literature produces a major bottleneck for interpreting and planning large-scale experiments. The ability to rapidly survey this literature is therefore a necessary step in both the design and the interpretation of such experiments. Moreover, automated text mining offers a yet untapped opportunity to integrate many fragments of information, gathered by researchers from multiple fields of expertise, into a complete picture exposing the interrelated roles of genes, proteins and chemical reactions in cells and organisms.

For these reasons there is a surge of interest in mining the biomedical literature,[1–11] (a comprehensive, growing list is available on the Biomedical Literature Mining Publication (BLIMP) website[12]), ranging from relatively modest tasks such as finding reported gene locations on chromosomes[7] to more ambitious attempts to construct putative gene networks based on gene-name co-occurrence within articles.[8] As the literature covers all aspects of biology, chemistry and medicine, theoretically there is almost no limit to the types of information that may be recovered through careful and exhaustive mining.

Regardless of the explicit purpose, there are several hurdles to overcome when looking for information in the biomedical literature. The sheer *number* of available articles (eg over 15,000,000 abstracts currently in PubMed,[13] and this number grows by the hour), makes it very difficult to find all and only the documents relevant to a specific need.

Hagit Shatkay,
School of Computing,
Queen's University,
Kingston, Ontario, K7L 3N6,
Canada

Tel: +1 613 533 6426
Fax: +1 613 533 6513
E-mail: shatkay@cs.queensu.ca

**Characteristics of biomedical text**

The *ambiguity* in both the English language and the biomedical jargon causes search engines to miss relevant papers and retrieve irrelevant ones.

Yet another issue is the inherent difference between the scientific literature and the text collections typically searched by current text-handling tools. Much of the work on text mining is aimed at (and tested on) articles such as news reports, typically written by professional writers whose main goal is to clearly convey a story to the average reader. In contrast, scientific documents are written by scientists whose first language may often not be English, whose main focus is research rather than report-writing, and whose target audience is a relatively small group of fellow scientists, all familiar with the same domain-specific jargon. Scientific articles thus often use non-standard terms and structures, and include material that may not directly pertain to – or may even *contradict* – the paper's main point. Finally, the purpose of mining the literature is not always crisp and clear, making it difficult to test and evaluate the merit of proposed biomedical text mining solutions. All these factors add a level of complexity, leading to a relatively poor performance of standard automated tools (see for instance Hersh[14] and Hirschman *et al.*[15]).

**Automated text mining comprises several research areas**

The automated handling of text is an active research area, spanning several disciplines, including the broad field of natural language processing (NLP),[16–18] the more specific domain within NLP of information extraction[19–21] and the area of information retrieval.[22–24] While NLP and information extraction are concerned with analysis of language and the mining of information within a given paper, information retrieval is concerned with the high-level task of obtaining the documents that may contain the information. All these techniques are being applied to a variety of tasks related to biomedical text mining. For example, in this issue there is a discussion of the use of natural language processing and ontologies in biomedicine[25] as well as of interaction network extraction,[26] which is an application of information extraction.

**Information retrieval is concerned with finding relevant documents**

This review focuses on the application of *information retrieval* to biomedical text. Information retrieval is a necessary first step towards text mining. It is the process of deciding which documents may contain relevant information, and to which of them other text-mining techniques should be applied. Often, information retrieval is used in and of itself, as exemplified by PubMed[13] – arguably the most widely used biomedical information retrieval tool. The following sections provide a survey of basic concepts and methods in information retrieval, discuss the way they are applied in the biomedical domain, and demonstrate the use of information retrieval in non-conventional ways toward obtaining facts about relationships among genes.

## INFORMATION RETRIEVAL: THE BASICS

Information retrieval is concerned with identifying, within a large document collection, a subset of documents whose content is most relevant to a user's need. More precisely, given a large database of documents, and a specific information need – usually expressed as a *query* by the user – the goal of information retrieval is to find the documents in the database that satisfy the information need. Naturally, the task has to be performed accurately and efficiently.

### Boolean queries and index structures

A simple and common way to express an information need is through a *Boolean* query. The user provides a term (eg OLE1), or a Boolean term-combination (eg OLE1 *and* lipid). The result produced by a retrieval system is the set of *all* the documents in the database satisfying the query constraints, eg containing both the query terms OLE1 and lipid. This query paradigm is used by the biomedical literature database PubMed, and by many other text search engines. It is supported

**Boolean queries retrieve documents containing specific words**

by an index covering all the terms in the whole database of documents. Each *term* may be a single word (eg 'blood') or a phrase (eg 'blood pressure'). It is common practice to omit from the index terms that are frequent and non-content-bearing, such as prepositions. These terms are referred to as 'stop words', and are usually viewed as delimiters when processing text. The index structure contains a sorted list of terms, and holds, for each term, a reference to all the documents in the database that contain it, as demonstrated in Figure 1. Further information on this topic is available in books concerning information management and access (eg Witten *et al.*[24]).

The simple form of Boolean query, which is efficiently implemented over large databases, suffers several limitations:

- The number of retrieved documents is typically *prohibitively large*.
- A substantial part of the retrieved documents is *irrelevant* to the user's information need.
- Many relevant documents *may not be retrieved*.

**Similarity queries and the vector model provide flexibility**

The second problem above stems from *polysemy*: a word may have multiple meanings in different contexts. For instance, the term '*arm*' may denote, among others, a limb, a part of the chromosome or a *Drosophila* gene (short for *armadillo*). On the other hand, the third problem stems from *synonymy*: a single concept is discussed in various abstracts under different names.

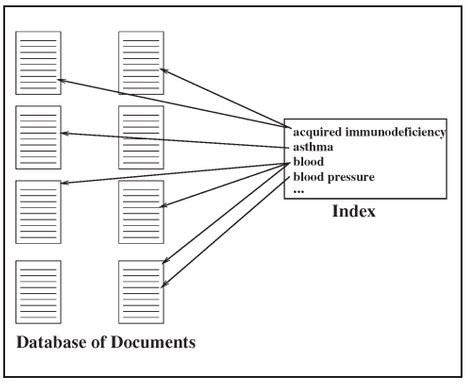


**Figure 1:** An index relating terms to the documents in which they occur

## Similarity queries and the vector model

A broadly used alternative to the Boolean query is the *similarity query*, which is typically based on the *vector-space* model, discussed throughout this section. Under this setting, documents are viewed as (algebraic) vectors over terms, as we formally define below. A query, $q$, may consist of many terms, and even comprise a complete document. It too is viewed as a body of text, rather than merely as a search-terms combination and is represented as a vector as well. The retrieval task reduces to searching the database for document-vectors that are *most similar* to the query-vector. Various similarity measures over documents have been devised and used.[23,27]

To explicitly define the vector model, we refer to the large set of documents to which retrieval is applied as the *database*, and denote it as *DB*. The *vocabulary* of the database is the set of terms occurring within *DB*'s documents. Let $M$ be the number of distinct terms $\{t_1, \ldots, t_M\}$ in this vocabulary. A term may be a single word or a combination such as 'acquired immunodeficiency syndrome'. A *document*, $d$, in the database is represented as an $M$-dimensional vector:

$$d = \langle w_{d_1}, w_{d_2}, \ldots, w_{d_M} \rangle,$$

where $w_{d_i}$ is a weight representing the occurrence or the significance of the term $t_i$ within the document $d$. The choice of term-weights can significantly influence the results of a similarity search, and there are many ways to calculate the weights.

For instance, the weight can be binary, either 1 or 0, corresponding to the presence or absence of a term in the document. While this representation is straightforward, it does not account for various properties of documents and terms that may improve retrieval. One simple extension uses the number of times the

term occurs within the document as the weight. An over-represented query-term within a document indicates that the document is likely to be relevant to the query.

**Term weighting schemes: TFX IDF and variations**

Yet another consideration is the distribution of query terms among documents. If one query-term frequently occurs in many documents, while another is rarely used, documents containing the rare term are likely to be more relevant to the query than documents containing the frequent one. These intuitions are combined and formalised through a family of weighting schemes commonly known as TF × IDF. The acronym stands for 'Term Frequency × Inverse Document Frequency'. Under this general scheme, the weight for a term is a product between the term frequency within the document and another number that is inversely proportional to the number of documents containing the term. Further discussion of weighting schemes is available in the extensive literature on information retrieval, eg Salton[23] and Witten *et al.*[24] In the biomedical context, Wilbur and Yang[28] study weighting schemes pertaining to retrieval from the biomedical literature.

**High weight is assigned to significant terms, which helps in finding the documents containing the most significant query terms**

Using the vector-space representation, we can apply a vector-similarity measure and assess similarity between pairs of documents as well as between a query and each document in the database. A similarity measure that is widely used in information retrieval is the *cosine coefficient*, which denotes the cosine of the angle between the two vectors, as illustrated in Figure 2. This measure normalises the vectors by their respective length, compensating for the difference in the number of terms between the typically short queries and the longer documents.

Formally, the cosine coefficient between two $n$-dimensional vectors, $V_1,\ V_2$, whose respective lengths (norms) are $\|V_1\|,\ \|V_2\|$, is defined as:

$$\cos(V_1,\ V_2) \stackrel{\text{def}}{=} \frac{\sum_{j=1}^{n} v_1^j \cdot v_2^j}{\|V_1\| \cdot \|V_2\|}.$$

**Relevance and pseudo-relevance feedback guide the search towards relevant documents**

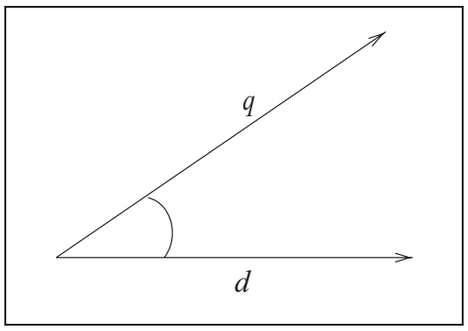


**Figure 2:** The angle between two vectors, *q* and *d*

## Other similarity-based approaches

Other approaches based on the vector-space model also aim to reduce the dependency of the retrieved documents on the particular choice of query terms, and effectively improve retrieval. One way to do this is through the re-weighting of query terms, where terms occurring within relevant documents receive a higher weight than those occurring in irrelevant ones. This process is called *relevance feedback*, and a re-weighting scheme introduced by Rocchio[29] is often used. Typically it is impractical to obtain actual relevance-judgment about the retrieved documents. Therefore, *pseudo* relevance feedback is used instead. Under this scheme (introduced by Buckley *et al.*[30]), given a current query, $Q_{\text{old}}$, the documents that rank high with respect to it are considered *relevant* while those ranking low are considered *irrelevant*. If the weight for term $t$ in $Q_{\text{old}}$ was $t_{Q_{\text{old}}}$, the new weight, $t_{Q_{\text{new}}}$, in the updated query is calculated as:

$$t_{Q_{\text{new}}} = A \cdot t_{Q_{\text{old}}} + B \cdot \text{avg\_t\_weight\_in\_rel\_docs} - C \cdot \text{avg\_t\_weight\_in\_nonrel\_docs}.$$

Here the parameters $A$, $B$ and $C$ measure the relative importance attributed to the original query $Q_{\text{old}}$ and to the relevant and irrelevant documents respectively,

avg_t_weight_in_rel_docs is the average weight of the term $t$ in the documents ranked as relevant to $Q_{old}$, and avg_t_weight_in_nonrel_docs is the average weight of the term $t$ in the irrelevant ones.

**Probalistic retrieval models**

Another way to relax the dependency between retrieval results and the explicit query terms is the use of *probabilistic models*.[27] Rather than require that a set (or a subset) of query terms occur in a document, the retrieval task is viewed as that of finding documents that with *high probability* satisfy the need represented by the query. Van Rijsbergen's work is one of the earliest in this direction.[31] More recent work include Ponte and Croft's language model,[32] and work on probabilistic latent semantics by Hofmann and others (eg Hofmann[33]). Another related method, in the biomedical context, about probabilistic themes,[34] was primarily applied to finding relationships among genes, as discussed later.

A different approach is *latent semantics analysis* introduced by Dumais *et al.*[35–38] Two main ideas underlie this method:

- There is an abstract semantics that explicit terms attempt to convey. Different terms may convey the same concept (synonymy), while a single term may denote different concepts (polysemy). While words are overtly present in a document, their semantics is not explicitly stated and is therefore *latent*.

- A collection of documents, each represented by an $M$-dimensional vector, can be viewed as a matrix. As such, algebraic operators can be applied to it. One particular operator, namely *singular value decomposition*, can be used to identify and extract the 'significant components', known as *singular values*, of the matrix.

**Machine learning methods are used for both hard and soft text categorisation**

By combining these two ideas, each of the $k$ large singular values of the matrix representing a document collection is viewed as a surrogate for a class of terms with a common *hidden* semantics. Both queries and documents are transformed and expressed as vectors over these singular values rather than as vectors over $M$ terms, and the similarity measure is applied to these transformed vectors, whose dimensionality is lower than that of the original term-space. Additional work in this direction has been carried out by several other research groups (eg Jiang and Littman[39] and Papadimitriou *et al.*[40]).

## Text categorisation

A task often addressed by information retrieval systems is that of *text categorisation*. This is the labelling of text by category-tags from a predefined set of categories. There are two main approaches to categorisation. One is the knowledge engineering approach[41,42] where the user manually defines a set of rules to encode expert knowledge regarding the correct categorisation of documents. The main drawback of this approach is the *knowledge acquisition bottleneck*. The rules must be manually defined by a knowledge engineer interviewing a domain expert. Any modification to the categories requires further intervention by the knowledge engineer.

The other is the machine learning (ML) approach,[43–53] where a text classifier is viewed as a function learnt by an inductive process, from a *training set* of example documents, already classified into a predefined set of categories. (As indicated in some of the references, a variation of the aforementioned Rocchio method is also applied in text categorisation.) ML-based classification is partitioned into two types: *hard* and *soft* classification. Under *hard* classification a document is strictly assigned to a single category, (eg Lewis *et al.*,[48] Buckley *et al.*[54] and Joachims[55]). In contrast, *soft* classification entails a *ranking* by relevance of the categories for each document. Under this approach, the classifier returns a number between 0 and 1 (called the categorisation status value, CSV), which represents the strength of evidence or the

probability that the document belongs to a certain category. Documents can then be ranked with respect to each category according to their CSV. (See Yang[56] for discussion and further references.)

**Supervised categorisation uses a training set and is known as classification. Unsupervised categorisation is known as clustering**

One final distinction made within machine learning categorisation is between *supervised* categorisation, known as *classification*, and *unsupervised* categorisation, known as *clustering*. Classification was discussed above, where a set of tags as well as a set of tagged training examples are given, and the learning task is to generate a classifier that could correctly assign tags to yet-unseen data items. In contrast, clustering is the partitioning of examples into coherent sets without the provision of predefined tags or training examples. In this case, the goal is to produce subsets (clusters), such that documents within a cluster are similar to each other according to some criteria, while documents contained in different clusters are dissimilar. Popular clustering methods include hierarchical clustering, *k*-means clustering,[57] and probabilistic approaches such as expectation maximisation.[58]

## Evaluation

**Evaluation measures are based on the ideas of specificity (precision) and sensitivity (recall)**

When developing a text-analysis tool, it is critical to know how reliable the results are likely to be. While we can neither anticipate all the articles we may encounter, nor predict performance in all cases, it is useful to evaluate the merit of a text-analysis tool by comparing its performance with that of other candidate techniques, with respect to a fixed gold standard. Such an evaluation requires two components:

- A corpus of annotated, tagged or categorised text items for a gold standard.
- A metric to measure the system's performance with respect to the gold standard.

Common measures for evaluating performance of information retrieval systems are *recall* and *precision*.[24,45,59] Recall ($R$) denotes the proportion of relevant articles retrieved by the system with respect to all the relevant articles in the data set that should have been retrieved (a notion similar to sensitivity). Precision ($P$) denotes the proportion of truly relevant articles among all the articles that were retrieved (similar to specificity). Other measures, such as the $F$-score[22] combine the two. In its simple form it is expressed as:

$$F = \frac{2PR}{P + R}.$$

For a thorough discussion of evaluation measures see Witten *et al.*,[24] Lewis[45] and Yang.[59]

A comparison of tools or methods requires an agreed upon corpus of reference, reflecting some true domain, with respect to which performance is measured. Several standard text collections, as well as standardised retrieval/extraction tasks have been devised – mostly during the past decade – especially for supporting development and evaluation of text-processing systems. Examples include the Reuters set classified into thematic categories,[60] and the OHSUMED collection of biomedical abstracts,[61] annotated with medical subject headings (MeSH) and tagged by relevance judgments with respect to specific queries. Both of these collections are used for evaluation in text-categorisation research.

A forum for standardised evaluation of retrieval methods is *TREC*, the Text Retrieval Conference,[62] sponsored by the National Institutes of Standards and Technology (NIST) and by DARPA. Each year it offers several tracks, each of which specifies data sets and tasks to be performed on them. Very recently a new track, *TREC Genomics*, concerned with the retrieval of genomic data from the literature and from other sources, has been formed.[14] The TREC Genomics effort, as well as another recent evaluation initiative, BioCreative, are discussed in another paper within this issue.[63]

## INFORMATION RETRIEVAL IN BIOMEDICAL INFORMATICS

Large-scale experimental methods, allowing analysis of genes and proteins from a whole genome, provide the first step towards understanding intricate cellular processes at the molecular level. While experiments are planned and carried out, both their informed planning and the interpretation of their results rely heavily on the ability to put new observations in the context of existing knowledge and of previous hypotheses. This type of information can often be found in the published literature. However, the conventional method for finding it has been for individuals to search through the literature, paper by paper and gene by gene. A tedious task for even a few genes, and almost impossible on a genomic scale.

**Biomedical information extraction looks for explicit statements about entities and relationships within the text**

To improve the effectiveness, efficiency and accuracy of the navigation through the literature, partial-automation of literature scanning is pursued in two main directions. First, much work focuses on *information extraction* from biomedical text. This includes the identification of named entities, such as genes and proteins,[9,15,64,65] and of relations such as location of genes on chromosomes,[2] association between a gene and a disorder or between proteins and sub-cellular organelles,[4,66] interaction among proteins[36,67–69] and many others.[5,70–75] These areas of research typically rely on lexical and ontological resources, and on techniques from natural language processing to identify facts, entities and certain structures within papers and sentences. These topics are addressed in several articles within this issue and the next.[25,26,76] The second direction addresses literature mining at a coarser granularity, namely that of finding, within a large database of articles, all and only the documents that contain relevant information, without extracting explicit facts from within the text. This approach is anchored in *information retrieval*, and is discussed throughout this section. The high availability and accessibility of abstracts (primarily through MEDLINE and PubMed, from the National Library of Medicine[13]), coupled with the limited access to full-text, accounts for the trend in most current biomedical text mining work, as described below, to focus on abstracts rather than on full-text articles.

### Boolean search methods and their extensions

**In its most basic form, biomedical information retrieval is exemplified by PubMed**

The most extensive and widely used information retrieval tool in the biomedical domain is the *PubMed* database and search-engine.[13] It contains over 15,000,000 scientific abstracts (mostly from MEDLINE, maintained by the National Library of Medicine, but also from other sources), and is accessed daily by millions of users throughout the world. For instance, during March 2005 alone over 68,000,000 PubMed searches were performed.[77]

A typical literature search within PubMed starts with a *Boolean* query. The user provides a term or a Boolean term-combination. The result is the set of *all* the abstracts in PubMed satisfying the query constraints, as discussed in the section on 'Boolean queries and index structures', above. We note that the lack of uniformity in nomenclature used by authors aggravates the problem of synonymy. For instance, a search for abstracts about the gene *AGP1* may not retrieve abstracts discussing this same gene under another name (eg *YCC5*). Still, if the user identifies a relevant document among those returned by the initial Boolean search, PubMed does offer a similarity-based tool (see the section 'Similarity queries and the vector model'), known as *neighboring*,[78] to access documents similar to the relevant one.

While PubMed is an indispensable resource, its size and breadth can make it difficult for researchers working on a specific organism or gene to obtain exactly the information they seek. In particular, as PubMed stores and searches only abstracts rather than full-text documents, it cannot locate information

that appears in the full text alone. Tools that provide access to small, organism- or topic-specific subsets of documents from PubMed, possibly including the full-text, have recently been suggested. As these tools are serving a smaller community and search over a smaller database, they can offer some enhanced functionality for the community they serve. An example is the Textpresso system,[79] which focuses on *C. elegans* and contains about 6,500 full papers and 20,000 curated abstracts. The system uses information extraction techniques to identify entities of interest (such as allele, process, function) based on defined ontologies, and provides the means to extend Boolean queries to match the specific types of entities.

Both PubMed and smaller literature databases such as Textpresso, primarily enable searching for information about 'one-gene-at-a-time', which is not suitable for large-scale literature mining. Moreover, PubMed is not primarily intended, and cannot be used 'as-is', for finding or explaining, on a large-scale, relationships among genes or other biological entities. Information extraction methods, as demonstrated in Textpresso, can be used to scan the documents retrieved by PubMed and search for facts. However, when there are many such documents, which are not all relevant to the focus of interest, information extraction can be hindered by exactly the same problems that caused inaccurate retrieval in the first place. For example, entities that may seem like gene names may not be genes at all, and much time will be spent scanning through documents that contain no relevant information whatsoever. We also note that information extraction aims to find in the literature information that is readily stated there. It does not aim to discover or deduce new relationships or facts that are not explicitly stated.

Thus, one important current goal of information retrieval is to reduce the size of the proverbial haystack, without requiring much human curation, to the point that a small enough set of relevant documents is available for information extraction methods to effectively and accurately obtain the desired facts. Another goal is to devise methods, based on information retrieval, to directly aid in the discovery of new facts and relationships within large corpora of documents. We next discuss ways in which IR is used in the context of discovery.

## Information retrieval as a basis for discovery in corpora

Early work on discovery of novel facts in the medical literature, predating the genomic era, was introduced by Swanson.[80–82] His method relies on 'transitive' relations, ie indirect links among entities, as clues for yet-unknown relationships. For instance, Swanson identified literature reports about *fish oil* causing *reduction in blood viscosity* and *decrease in platelet aggregability*. He also identified a different body of literature discussing both of these symptoms as characteristic of *Raynaud's syndrome*. Based on these two sets of seemingly unrelated reports, he established the *hitherto unknown connection*, namely that fish oil can treat Raynaud's syndrome. This connection is illustrated in Figure 3.

**Early work on discovery through retrieval was done by Swanson, in the medical domain. It was based on well-chosen Boolean queries and examination of overlap in the results**

Put simplistically, the discovery method for relating entities A and B (where A is fish oil, B is Raynaud's syndrome in the above example) consists of the retrieval of all the documents containing term A and all those containing term B. If these two bodies of literature do not overlap, try to find concepts C occurring in both of them. The concepts C can indicate yet-undiscovered relationships between concepts A and B. This line of reasoning was developed and further automated by Weeber *et al.*,[83] and Srinivasan and Libbus,[84,85] as well as by Wren.[86]

In the context of large-scale genomics, methods to support biomedical analysis based on information retrieval have also been introduced and developed during the last few years. Shatkay *et al.*[6,87] introduced an information retrieval scheme to find functional relationships

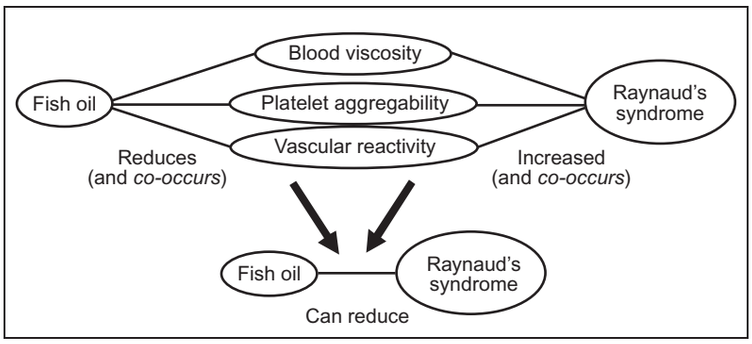


**Figure 3:** Swanson discovery of the link between fish oil and Raynaud's syndrome. The link was made by finding common terms, such as blood viscosity, in the literature discussing these two separate concepts

**In early work on genomic discovery through retrieval, thematic analysis is used for finding functional relationships among genes**

among genes. The method comprises three main ideas:

- Analysis of large-scale gene expression data (microarrays) can benefit from the automated introduction of text related to the genes.
- Text discussing a gene can be used as a 'surrogate' representation for the gene.
- Once genes are represented by their text documents and/or terms, this representation can be clustered to find genes with similar behaviours (similar themes), in ways reminiscent of the clustering methods applied to gene sequences or gene expression profiles.

Basing the method on information retrieval, rather than on extraction, reduces the dependency on gene nomenclature or on sentence structure. The approach relies on the fact that many individual genes and their function are already discussed in the literature. A database containing tens of thousands of PubMed abstracts, pertaining to a specific domain, is used. To find relationships among a large set of genes, each gene is mapped to a single abstract within the collection, discussing the gene's biological function. This abstract is treated as the gene's *representative*. A probabilistic theme-finding algorithm[34] finds a set of documents relevant to this abstract, and produces a *set of terms* summarising the thematic contents of the document set. Applying the algorithm to each gene-representing abstract produces for each gene a body of related literature (20–50 abstracts bearing a common *theme*), along with a list of terms that characterise its theme. An automated comparison of the abstract sets associated with the different genes is then used to derive relationships among them. This method was tested on 400 yeast genes over a database of about 40,000 PubMed abstracts; a study of the biological roles of yeast genes by Spellman *et al.*[88] was used as a gold standard. In addition, a thesaurus of biological function terms for yeast genes, built by a panel of four yeast experts, was used to quantitatively evaluate the list of characteristic terms. The results showed that for about 100 genes an informative representative abstract was found. For these, the related genes identified by the system typically shared the same biological function, and on average, three or four of the top five terms assigned by the system to each gene correctly indicated its function. A summary of the method is shown in Figure 4.

## Text categorisation for bioinformatics

In recent years several groups have applied clustering and classification methods to text in the context of bioinformatics. It is important to note that all the studies described here vary in goal and scope, and suffer from the lack of agreed-upon evaluation standards for testing their performance.

Renner and Aszódi[89] suggested a method for clustering protein annotations. The basic idea is that by clustering the annotations of proteins into groups one can gain insight into the common function that the proteins may have. The method is based on first grouping terms that occur in protein annotations into sets, according to their tendency to co-occur. A similarity measure among the annotations is then devised based on the proportion of terms in them that are in

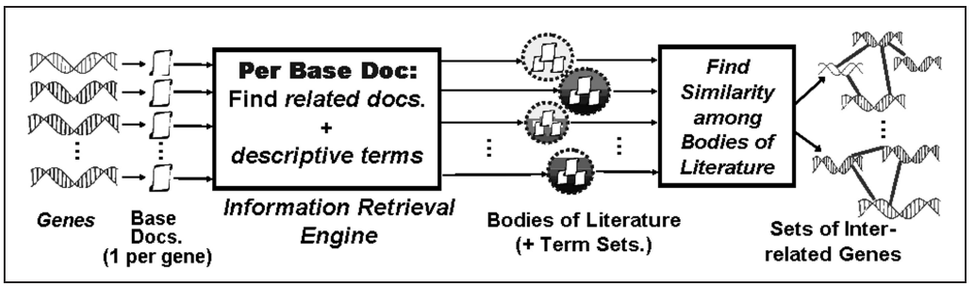


**Figure 4:** Finding functional relations among genes via thematic analysis of the literature. Each gene is represented as an abstract discussing the gene. A theme consisting of related documents and summary terms is built around each abstract. Genes with similar themes are grouped together

**Clustering of protein annotation text can be used to group together related proteins**

the same term-groups. Finally, the annotations are clustered using hierarchical clustering based on the suggested similarity measure. The method was tested by grouping 900 terms into 400 clusters. Clustering of documents based on the term groups was tested on two sets of fewer than 50 PubMed abstracts. The method was applied to the analysis of a set containing about 3,000 protein annotations.

Stephens *et al.*[90] deduce relationships among genes based on co-occurrence of their names (where the names are given in a thesaurus), which is a subject typically handled by information extraction (eg Jenssen *et al.*[8]). However, this work uses information retrieval techniques to identify co-occurrence. Documents are represented as weight vectors, where the genes mentioned in the documents are the terms. Looking at the transposed matrix, each gene is then viewed as a vector whose attributes reflect the documents that mention it. The association between genes is measured by the dot product between the two vectors representing them (which corresponds to the unnormalised cosine coefficient). This measure essentially quantifies the co-occurrence of the genes within documents. The method was applied to a set of about 5,000 PubMed abstracts.

Iliopoulos *et al.*[91] apply *k*-means clustering[57] to a relatively small set of PubMed abstracts (fewer than 2,000 documents) to obtain meaningful subsets, each discussing some common subject. The subjects are then represented through terms extracted by statistical analysis of term frequencies within the clusters. The effectiveness of each of the three methods above was demonstrated using a limited example, but a thorough quantitative evaluation was not reported.

Marcotte *et al.*[92] apply a naïve Bayes classifier[93] that relies on discriminating terms, to identify abstracts that discuss protein–protein interactions. The classifier was tested on a set of 325 PubMed abstracts, of which 70 discussed protein–protein interactions and 255 did not. The performance on this limited set was about 0.6 precision and recall at the break-even point.

While the above methods categorise a (relatively small) collection of abstracts into topical sets, others apply categorisation to directly enhance, through the use of text, specific applications that typically rely on biological data.

Using an idea similar to the one presented earlier[6] of viewing an abstract as a 'surrogate' or a representative for a gene, Stapley *et al.*[94] represent proteins by using the abstracts that mention them. They then train a support vector machine (SVM) classifier to distinguish among abstracts discussing proteins, based on the different subcellular locations of the proteins mentioned in the text. They propose this classifier as an aid for addressing the protein sorting problem, which is the task of determining the organelle within the cell to which the

protein belongs. (Note that the same task was tackled through information extraction, looking for statements about subcellular localisation in the literature, by Craven and Kumlien.[4]) Another research on text-based classification of proteins into subcellular locations was performed by Nair and Rost.[95] In this case proteins were represented by keywords obtained from the protein annotation in SwissProt, and classification was applied to the keyword-vector representation of the proteins, rather than to PubMed abstracts. While the approach is interesting, the results reported in both of these papers[94,95] did not improve upon the state-of-the-art in protein subcellular localisation (eg Emanuelsson *et al.*[96] and Park and Kanehisa[97]).

In a recent application,[98] clustering of genes was performed using latent semantics analysis (LSA)[37,38] of the abstracts representing the genes. The idea of clustering abstracts as surrogates for genes is similar to that discussed above. In this case, however, abstracts discussing each gene were concatenated to create a single document representing the gene, and LSA was used to reduce the dimensionality of the representation of these documents and to cluster together genes with similar topics. This study was done on a very small scale (50 genes and gene-documents), and was used to reconstruct the 18-gene Reelin signalling pathway in mouse.

## Information retrieval in integrative applications

**Information retrieval may enhance other types of biological data analysis**

Several lines of work have recently incorporated methods from information retrieval into the analysis of other types of data, to enhance the performance of experimental data analysis.

The first such work we are aware of is in the realm of protein homology, performed by Chang *et al.*[99] In this work, text that accompanies protein sequences is used to support homology search. As PSI-BLAST is applied to the protein sequences to detect homology among proteins, it is augmented with the cosine similarity measure which is applied to the accompanying text. While this attempt was a novel integration of text and protein sequence data, the results did not suggest a significant improvement over simple protein homology.

Recent research by Glenisson *et al.*,[100,101] which directly continues the line of work on large-scale gene expression analysis using text,[6] integrates clustering of documents into the process of clustering gene expression data. It suggests that the integration of text clustering with the clustering of the expression of the genes discussed in the text, produces more coherent and stable clusters than those produced by expression alone. Work by Raychaudhuri *et al.*[102] also suggests that text clustering results in coherent clustering of the genes represented by the text.

Another form of an integrative method combines information extraction and information retrieval. Donaldson *et al.*[103] introduced their PreBind/Textomy system, in which they combine the two approaches to assist in recovering protein–protein interactions from the literature. On the information retrieval step, an SVM classifier is trained to distinguish between PubMed abstracts that discuss protein–protein interaction and abstracts that do not. The classifier is then used to identify and retrieve the abstracts that are relevant to protein–protein interaction. Once they are retrieved, information extraction is applied to identify interaction facts within the text. The SVM is used in this phase again to find sentences containing the interaction information. From each such sentence, protein names are extracted (based on a list of protein names and synonyms), as candidates for protein–protein interaction. Simple co-occurrence of protein names within the complete abstract is also an indicator for possible interaction between the proteins. The system serves as a curation aid for the BIND database.[104] The putative interactions found are not meant to be automatically placed in an interaction

network. Rather, the BIND curators examine and validate them by reading the related text. The retrieval system was trained and tested, on a set of about 1,100 expert-judged abstracts of which about 700 discuss interactions and 400 do not. The success rate reported is 92 per cent in both precision and recall for identifying abstracts discussing interaction. The extraction of actual protein–protein interactions was tested by comparison to a list of about 1,380 human-curated protein–protein interactions in yeast (restricted to interactions reported in the literature). About 60 per cent of these interactions were successfully recovered from the abstracts that were classified as discussing interactions. The higher accuracy of the early classification phase suggests that a major advantage of the system lies in the retrieval step, which identifies documents relevant to protein–protein interaction.

We note that almost all of the work surveyed here was based on abstracts rather than on full-text articles, owing to the high availability of abstracts. There is an ongoing effort to obtain full-text documents for biomedical text mining, as is evident in the recent TREC Genomics categorisation task,[14] and in the launching of several open access journals. Underlying such efforts is the assumption that there is much information in the full text, where text-mining methods could demonstrate their true utility. While the full text enables access to much more information, including complete experimental reports, observations, facts and hypotheses, it is also longer, less concise and has more room for ambiguity than the typical text in abstracts. These latter factors may increase processing time and space, and place a heavier burden on text analysis tools. Thus, the pros and cons of mining full-text biomedical articles are yet to be studied.

## CONCLUSIONS

The abundance of biomedical literature motivates an intensive pursuit for effective text-mining tools. Such tools are expected to help uncover the information present in the large and unstructured body of text, while addressing three main problems:

- The sheer magnitude of the available text collections.
- The ambiguity and non-uniformity of the nomenclature used in the context of genomics and proteomics.
- The linguistic complexity of the scientific documents, stemming from the diversity in expertise, style and native language of the authors.

In general, information retrieval provides the means for a coarse-grain search for relevant documents. While it is not intended to extract a tidy fact statement, it can produce a relatively small set of choice documents, thus restricting the search-space within which the facts of interest can be found. This focused set of documents can also provide the relevant literature needed for analysing and explaining experimental results (on which other automated mining systems may operate). Moreover, we have shown that a non-traditional use of information retrieval can actually provide an effective way for detecting specific putative relationships among genes.[6,100] Since information retrieval does not look for explicitly stated facts within the literature, it has the potential to *foreshadow yet undiscovered facts*. A clear advantage of the information retrieval approach is its relative independence of specific natural language usage and nomenclature issues, as it does not search for explicit gene names or statements about their relationships. The latter is a major feature given the complex and incomplete nomenclature of the biomedical domain.

Perhaps the most important point demonstrated in the last section is the need for uniform standards by which system performance can be measured. The construction of gold standards and procedures for evaluating the utility of biomedical literature-mining tools is a

high-priority task. Efforts in this direction are discussed in another paper within this issue.[63]

As literature mining challenges in the context of bioinformatics vary widely in scope, data sources and ultimate goals, no single tool can currently perform all the required tasks. However, a combination of methods is likely to address many of the problems. Several such combinations of data and methods were discussed in the section on integrative applications.

To successfully mine the biomedical literature, it is important to realise the merits and the limitations of the different literature-mining methods. Moreover, it is essential to coherently state the actual biomedical problems we expect to address by using such methods. The ad-hoc retrieval task in TREC genomics is currently moving in the direction of problem-specific retrieval. Moving away from generic, all-purpose biomedical text mining solutions, and focusing the efforts on specific needs, is likely to expedite progress both in biomedical text mining and in large-scale biology.

**Acknowledgment**
Hagit Shatkay's work is partially funded by NSERC Discovery Grant #298292-04.

## *References*

1. Andrade, M. A. and Valencia, A. (1997), 'Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system', in 'Proceedings of the AAAI Conference on Intelligent Systems for Molecular Biology (ISMB)', AAAI Press, Menlo Park, CA, pp. 25–32.
2. Leek, T. R. (1997), 'Information extraction using hidden Markov models', Master's thesis, Department of Computer Science, University of California, San Diego.
3. Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T. (1998), 'Toward information extraction: identifying protein names from biological papers', in 'Proceedings of the 3rd Pacific Symposium on Biocomputing (PSB)', 4th–9th January, Hawaii, pp. 705–716.
4. Craven, M. and Kumlien, J. (1999), 'Constructing biological knowledge bases by extracting information from text sources', in 'Proceedings of the AAAI Conference on Intelligent Systems for Molecular Biology (ISMB)', AAAI Press, Menlo Park, CA, pp. 77–86.
5. 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 5th Pacific Symposium on Biocomputing (PSB)', 4th–9th January, Hawaii, pp. 514–525.
6. Shatkay, H., Edwards, S., Wilbur, W. J. and Boguski, M. (2000), 'Genes, themes and microarrays: Using information retrieval for large scale gene analysis', in 'Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)', AAAI Press, Menlo Park, CA, pp. 317–328.
7. Friedman, C., Kra, P., Yu, H., *et al.* (2001), 'GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles', in 'Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)', AAAI Press, Menlo Park, CA, pp. S74–S82.
8. Jenssen, T.-K., Laegreid, A., Komorowski, J. and Hovig, E. (2001), 'A literature network of human genes for high-throughput analysis of gene expression', *Nature Gen.*, Vol. 28, pp. 21–28.
9. Hirschman, L., Morgan, A. A. and Yeh, A. S. (2002), 'Rutabaga by any other name: Extracting biological names', *J. Biomed. Informatics*, Vol. 35(4), pp. 247–259.
10. Shatkay, H. and Feldman, R. (2003), 'Mining the biomedical literature in the genomic era: An overview', *J. Comput. Biol.* Vol. 10(6), pp. 821–855.
11. Hanisch, D., Fluck, J., Mevissen, H. T. and Zimmer, R. (2003), 'Playing biology's name game: Identifying protein names in scientific text', in 'Proceedings of the 8th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 403–411.
12. URL: http://blimp.cs.queensu.ca
13. PubMed (URL: http://www.ncbi.nlm.nih.gov/pubmed).
14. Hersh, W. (2003), 'TREC genomics track' (URL: http://ir.ohsu.edu/genomics).
15. Hirschman, L., Yeh, A., Blaschke, C. and Valencia, A. (2005), 'Overview of BioCreAtIvE: Critical assessment of information extraction for biology', *BMC Bioinformatics*, Vol. 6(S1).
16. Charniak, E. (1993), 'Statistical Language Learning', MIT Press, Cambridge, MA.
17. Allen, J. (1995), 'Natural Language Understanding', Benjamin Cummings, Redwood City, CA.
18. Manning, C. and Schütze, H. (1999),

'Foundations of Statistical Natural Language Processing', MIT Press, Cambridge, MA.

19. Cowie, J. and Lehnert, W. (1996), 'Information extraction', *Commun. ACM*, Vol. 39(1), pp. 80–91.
20. Cardie, C. (1997), 'Empirical methods in information extraction', *AI Magazing*, Vol. 18(4), pp. 65–80.
21. Grishman, R. (1997), 'Information extraction: techniques and challenges', in 'Proc. SCIE', 14th–18th July, Frascati, pp. 10–27.
22. van Rijsbergen, C. J. (1979), 'Information Retrieval', Butterworth, London.
23. Salton, G. (1989), 'Automatic Text Processing', Addison-Wesley, Reading, MA.
24. Witten, I. H., Moffat, A. and Bell, T. C. (1999), 'Managing Gigabytes, Compressing and Indexing Documents and Images', 2nd edn, Morgan-Kaufmann, San Francisco.
25. Spasic, I., Ananiadou, S., McNaught, J. and Kumar, A. (2005), 'Text mining and ontologies in biomedicine: Making sense of raw data', *Brief. Bioinformatics* (Special Issue on Text Mining), Vol. 6, pp. 239–251.
26. Skusa, A., Rüegg, A. and Köhler, J. (2005), 'Extraction of biological interaction networks from scientific literature', *Brief. Bioinformatics* (Special Issue on Text Mining), Vol. 6, pp. 263–276.
27. Sparck-Jones, K., Walker, S. and Robertson, S. (1998), 'A probabilistic model of information retrieval: Development and status', Technical Report TR446, University of Cambridge, Computer Laboratory (URL: http://www.ftp.cl.cam.ac.uk/ftp/papers/reports/#TR446).
28. Wilbur, W. J. and Yang, Y. (1996), 'An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology text', *Computers Biol. Med.*, Vol. 26(3), pp. 209–222.
29. Rocchio, J. J. (1971), 'Relevance feedback in information retrieval', in Salton, G., Ed., 'The SMART Retrieval System: Experiments in Automatic Document Processing', Prentice Hall, Englewood Cliffs, NJ, pp. 313–323.
30. Buckley, C., Salton, G., Allan, J. and Singhad, A. (1994), 'Automatic Query Expansion using SMART: TREC 3', in 'Proceedings of the Third Text Retrieval Conference (TREC-3)', NIST, Gaithersburg, MD, pp. 69–80.
31. van Rijsbergen, C. J. (1977), 'A theoretical basis for the use of co-occurrence data in information retrieval', *J. Documentation*, Vol. 33(2), pp. 106–119.
32. Ponte, J. M. and Croft, W. B. (1998), 'A language modeling approach to information retrieval', in 'Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR-98)', ACM Press, New York, pp. 275–281.
33. Hofmann, T. (1999), 'Probabilistic latent semantic indexing', in 'Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR-99)', ACM Press, Washington, DC, pp. 50–57.
34. Shatkay, H. and Wilbur, W. J. (2000), 'Finding themes in MEDLINE documents: Probabilistic similarity search', in 'Proceedings of the IEEE Conference on Advances in Digital Libraries', 2nd–7th June, Bethesda, MD, TX, pp. 183–192.
35. Dumais, S. T., Furnas, G. W., Landauer, T. K., *et al.* (1988), 'Using latent semantic analysis to improve access to textual information', in 'Proceedings of the Conference on Human Factors in Computing (CHI88)', ACM Press, Washington, DC, pp. 281–285.
36. Furnas, G. W., Deerwester, S., Dumais, S. T., *et al.* (1988), 'Information retrieval using a singular value decomposition model of latent semantic structure', in 'Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR-88)', ACM Press, New York, pp. 465–480.
37. Deerwester, S., Dumais, S. T., Furnas, G. W., *et al.* (1990), 'Indexing by latent semantic analysis', *J. Soc. Information Sci.*, Vol. 41(6), pp. 391–407.
38. Dumais, S. T. (1990), 'Enhancing performance in latent semantic (LSI) indexing', *Behavior Res. Methods, Instruments Computers*, Vol. 23(2), pp. 229–236.
39. Jiang, F. and Littman, M. (2000), 'Approximate dimension equalization in vector-based information retrieval', in 'Proceedings of International Conference on Machine Learning (ICML)', Morgan Kaufman, San Francisco, pp. 423–430.
40. Papadimitriou, C. H., Tamaki, H., Raghavan, P. and Vempala, S. (1998), 'Latent semantics indexing: a probabilistic analysis', in 'Proceedings of the Seventeenth ACM Symposium on Principles of Databases', ACM Press, New York, pp. 159–168.
41. Hayes, P. and Weinstein, S. (1990), 'CONSTRUE: A system for content-based indexing of a database of news stories', in 'Proceedings of the 2nd Annual Conference on Innovative Applications of Artificial Intelligence', AAAI Press, Menlo Park, CA.
42. Hayes, P. (1992), 'Intelligent high-volume processing using shallow, domain-specific techniques', in Jacobs, P., Ed., 'Text-based Intelligent Systems: Current Research and

Practice in Information Extraction and Retrieval', Lawrence Erlbaum, Hillsdale, NJ, pp. 227–242.

43. Cohen, W. W. and Singer, Y. (1999), 'Context-sensitive learning methods for text categorization', *ACM Trans. Information Syst.*, Vol. 17(2), pp. 141–173.

44. Riloff, E. and Lehnert, W. (1994), 'Information extraction as a basis for high-precision text classification', *ACM Trans Info. Systems*, Vol. 12(3), pp. 296–333.

45. Lewis, D. D. (1995), 'Evaluating and optimizing autonomous text classification systems', in 'Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR-95)', ACM Press, New York, pp. 246–254.

46. Vapnik, V. (1995), 'The Nature of Statistical Learning Theory', Springer, New York.

47. Larkey, L. S. and Croft, W. B. (1996), 'Combining classifiers in text categorization', in 'Proceedings of 19th ACM International Conference on Research and Development in Information Retrieval (SIGIR-96)', ACM Press, New York, pp. 289–297.

48. Lewis, D. D., Schapire, R. E., Callan, J. P. and Papka, R. (1996), 'Training algorithms for linear text classifiers', in 'Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (SIGIR-96)', ACM Press, New York, pp. 298–306.

49. Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M. (1998), 'Inductive learning algorithms and representations for text categorization', in 'Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM-98)', 3rd–7th November, Bethesdu, MD, pp. 148–155.

50. Joachims, T. (1998), 'Text categorization with support vector machines: Learning with many relevant features', in 'Proceedings of the European Conference on Machine Learning (ECML-98), Chemnitz, Germany, pp. 137–142.

51. McCallum, A. and Nigam, K. (1998), 'A comparison of event models for naive Bayes text classification', in 'Proceedings of the AAAI/ICML Workshop on Learning for Text Categorization', AAAI Press, Menlo Park, CA, pp. 41–48.

52. Yang, Y. and Liu, X. (1999), 'A re-examination of text categorization methods', in 'Proceedings of the 22nd ACM International Conference on Research and Development in Infromation Retrieval (SIGIR-99)', ACM Press, New York, pp. 42–49.

53. Sebastiani, F. (2002), 'Maching learning in automated text categorization', *ACM Comput. Surveys*, Vol. 34(1), pp. 1–47.

54. Buckley, C., Allan, J. and Salton, G. (1994), 'Automatic routing and ad-hoc retrieval using SMART: TREC 2', in 'Proceedings of the Second Text Retrieval Conference (TREC)', NIST, Gaithersburg, MD, pp. 45–56.

55. Joachims, T. (1997), 'A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization', in 'Proceedings of the International Conference on Machine Learning (ICML-97)', Nashville, TN, pp. 143–151.

56. Yang, Y. (2001), 'A study of thresholding strategies for text categoriztion', in 'Proceedings of the 24th ACM International Conference on Research and Development in Information Retrieval (SIGIR-2000)', ACM Press, New York, pp. 137–145.

57. Duda, R. O. and Hart, P. E. (1973), 'Unsupervised learning and clustering', in 'Pattern Classification and Scene Analysis', (Chapter 6), John Wiley and Sons. New York.

58. Cheeseman, P. and Stutz, J. (1996), 'Bayesian classification (AUTOCLASS): Theory and results', in Fayad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., Eds, 'Advances in Knowledge Discovery and Data Mining', AAAI Press, Menlo Park, CA, pp. 153–180.

59. Yang, Y. (1999), 'An evaluation of statistical approaches to text categorization', *Information Retrieval*, Vol. 1, pp. 69–90.

60. Lewis, D. D., 'Test Collections: Reuters-21578' (URL: http://www.daviddlewis.com/resources/testcollections/reuters21578).

61. Hersh, W., Buckley, C., Leone, T. J. and Hickam, D. (1994), 'OHSUMED: An interactive retrieval evaluation and new large test collection for research', in 'Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR-94)', 3rd–6th July, Dublin, pp. 192–201.

62. Voorhees, E. and Harman, D. K. (2001), Text REtrieval Conference (TREC) (URL: http://trec.nist.gov).

63. Hersh, W. (2005), 'Evaluation of biomedical text-mining systems: Lessons learned from information retrieval', *Brief. Bioinformatics*, in press.

64. Tanabe, L. and Wilbur, W. J. (2002), 'Tagging gene and protein names in full text articles', in 'Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain', 11th June, Philadelphia, PA, pp. 9–13.

65. Schwartz, A. and Hearst, M. (2003), 'Simple algorithm for identifying abbreviation

definitions in biomedical text', in 'Proceedings of the 8th Pacific Symposium on Biocomputing (PSB2003)', 3rd–7th January, Hawaii, pp. 451–462.

66. Ray, S. and Craven, M. (2001), 'Representing sentence structure in hidden Markov models for information extraction', in 'Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-01)', MIT Press, Cambridge, MA, pp. 177–210.

67. Blaschke, C., Andrade, M. A., Ouzounis, C. and Valencia, A. (1999), 'Automatic extraction of biological information from scientific text: Protein–protein interactions', in 'Proceedings of the AAAI Conference on Intelligent Systems for Molecular Biology (ISMB)', pp. 60–67, AAAI Press, Menlo Park, CA.

68. Temkin, J. and Gilder, M. (2003), 'Extraction of protein interaction information from unstructured text using a context-free grammar', *Bioinformatics*, Vol. 19(16), pp. 2046–2053.

69. Daraselia, N., Egorov, S. Yazhak, A., *et al.* (2004), 'Extracting human protein interactions from MEDLINE using a full-sentence parser', in 'Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics', 24th September, Pisa, pp. 11–17.

70. Tanabe, L., Scherf, U., Smith, L. H., *et al.* (1999), 'MedMiner: An Internet text-mining tool for biomedical information with application to gene expression profiling', *BioTechniques*, Vol. 27(6), pp. 1210–1217.

71. Park, J. C., Kim, H. S. and Kim, J. J. (2001), 'Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar', in 'Proceedings of the 6th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 396–407.

72. Yakushiji, A., Tateisi, Y. and Miyao, Y. (2001), 'Event extraction from biomedical papers using a full parser', in 'Proceedings of the 6th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 408–419.

73. Hahn, U., Romacker, M. and Schultz, S. (2002), 'Creating knowledge repositories from biomedical reports: The MEDSYNDIKATE text mining system', in 'Proceedings of the 7th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 338–349.

74. Pustejovsky, J., Castaño, J., Zhang, J., *et al.* (2002), 'Robust relational parsing over biomedical literature: Extracting inhibit relations', in 'Proceedings of the 7th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 362–373.

75. Shah, P., Perez-Iratxeta, C., Bork, P. and Andrade, M. A. (2003), 'Information extraction from full text scientific articles: Where are the keywords?', *BMC Bioinformatics*, Vol. 4(1), p. 20.

76. Kumar, A. (2005), 'What makes a gene name? Named entity recognition in the biomedical literature', *Brief. Bioinformatics*, accepted for publication.

77. URL: http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmeddata.htm

78. Wilbur, W. J. and Coffee, L. (1994), 'The effectiveness of document neighboring in search enhancement', *Information Proc. Manage.*, Vol. 30(2), pp. 253–266.

79. Müller, H.-M., Kenny, E. E. and Sternberg, P. W. (2004), 'Textpresso: An ontology-based information retrieval and extraction system for biological literature', *PLoS Biol.*, Vol. 2(11), pp. 1984–1998.

80. Swanson, D. R. (1986), 'Fish-oil, Raynaud's syndrome and undiscovered public knowledge', *Perspectives Biol. Med.*, Vol. 30(1), pp. 7–18.

81. Swanson, D. R. (1988), 'Migraine and magnesium: Eleven neglected connections', *Perspectives Biol. Med.*, Vol 31(4), pp. 526–557.

82. Swanson, D. R. (1990), 'Somatomedin C and arginine: Implicit connections between mutually isolated literatures', *Perspectives in Biol. Med.*, Vol. 33(2), pp. 157–186.

83. Weeber, M., Klein H., de Jong-van den Berg, L. T. W. and Vos, R. (2001), 'Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries', *J. Amer. Soc. Information Sci.*, Vol. 52(7), pp. 548–557.

84. Srinivasan, P. and Libbus, B. (2004), 'Mining MEDLINE for implicit links between dietary substances and diseases', in 'Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB), AAAI Press, Menlo Park, CA, pp. 290–296.

85. Srinivasan, P. (2004), 'Text mining: Generating hypotheses from MEDLINE', *J. Amer. Soc. Information Sci.*, Vol. 55(5), pp. 396–413.

86. Wren, J. D. (2004), 'Extending the mutual information measure to rank inferred literature relationships', *BMC (BioMed Central) Bioinformatics*, Vol. 5(1), p. 145.

87. Shatkay, H., Edwards, S. and Boguski, M. (2002), 'Information retrieval meets gene analysis', *IEEE Intelligent Systems* (Special Issue on Intelligent Systems in Biology), Vol. 17(2), pp. 45–53.

88. Spellman, P. T., Sherlock, G., Zhang, M. Q., *et al.* (1998), 'Comprehensive identification of cell cycle-regulated genes of the yeast

*Saccharomyces cerevisiae* by microarray hybridization', *Mol. Biol. Cell*, Vol. 9, pp. 3273–3297.

89. Renner, A. and Aszódi, A. (2000), 'High-throughput functional annotation of novel gene products using document clustering', in 'Proceedings of the 5th Pacific Symposium on Biocomputing (PSB)', 4th–9th January, Hawaii, pp. 54–68.

90. Stephens, M., Palakal, M., Mukhopadhyay, S., *et al.* (2001), 'Detecting gene relations from Medline abstracts', in 'Proceedings of the 6th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 483–496.

91. Iliopoulos, I., Enright, A. J. and Ouzounis, C. (2001), 'TEXTQUEST: Document clustering of MEDLINE abstracts for concept discovery in molecular biology', in 'Proceedings of the 6th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 384–395.

92. Marcotte, E. M., Xenarios, I. and Eisenberg, D. (2001), 'Mining literature for protein–protein interactions', *Bioinformatics*, Vol. 17(4), pp. 359–363.

93. Russell, S. and Norvig, P. (2002), 'Artificial Intelligence: A Modern Approach', Prentice Hall, Englewood Cliffs, NJ.

94. Stapley, B. J., Kelley, L. A. and Sternberg, M. J. E. (2002), 'Predicting the subcellular location of proteins from text using support vector machines', in 'Proceedings of the 7th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 374–385.

95. Nair, R. and Rost, B. (2002), 'Inferring sub-cellular localization through automated lexical analysis', *Bioinformatics*, Vol. 18 (Suppl. 1), pp. S78–S86.

96. Emanuelsson, O., Nielsen, H., Brunak, G. and von Heijne, G. (2000), 'Predicting subcellular localization of proteins based on their N-terminal amino acid sequence', *J. Mol. Biol.*, Vol. 300, pp. 1005–1016.

97. Park, K.-J. and Kanehisa, M. (2003), 'Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs', *Bioinformatics*, Vol. 19(13), pp. 1656–1663.

98. Homayouni, R., Heinrich, K., Wei, L. and Berry, M. (2005), 'Gene clustering by latent semantics indexing of MEDLINE abstracts', *Bioinformatics*, Vol. 21(1), pp. 104–115.

99. Chang, J. T., Raychaudhuri, S. and Altman, R. B. (2001), 'Including biological literature improves homology search', in 'Proceedings of the 6th Pacific Symposium on Biocomputing (PSB)', 3rd–7th January, Hawaii, pp. 374–383.

100. Glenisson, P. (2004), 'Integrating scientific literature with large scale gene expression analysis', PhD thesis, Katholieke Universiteit Leuven, Belgium.

101. Glenisson, P., Mathys, J. and Moor, B. D. (2003), 'Meta-clustering of gene expression data and literature-extracted information', *ACM SIG KDD Explorations* (Special Issue on Microarray Data Mining), Vol. 5(2), pp. 101–112.

102. Raychaudhuri, S., Schütze, H. and Altman, R. B. (2002), 'Using text analysis to identify functionally coherent gene groups', *Genome Res.*, Vol. 12(10), pp. 1582–1590.

103. Donaldson, I., Martin, J., de Bruijn, B., *et al.* (2003), 'PreBind and textomy – mining the biomedical literature for protein–protein interactions using a support vector machine', *BMC (BioMed Central) Bioinformatics*, Vol. 4(11) (URL: http://www.biomedcentral.com/1471-2105/4/11).

104. Bader, G. D., Betel, D. and Hogue, C. W. V. (2003), 'BIND: The Biomolecular Interaction Network', *Nucleic Acids Res.*, Vol. 31(1), pp. 248–250 (URL: http://www.bind.ca).