A survey of current work in biomedical text mining

Aaron M. Cohen and William R. Hersh

Date received (in revised form): 25th October 2004

Aaron M. Cohen, MD
is a postdoctoral fellow in the medical informatics programme at OHSU. Dr Cohen works in the area of text mining, focusing on issues and applications important to biomedical researchers. He was chairman of the W3C working group that produced version 2 of the Synchronized Multimedia Integration Language (SMIL 2.0).

William Hersh, MD
is Professor and Chair of the Department of Medical Informatics & Clinical Epidemiology in the School of Medicine at Oregon Health & Science University (OHSU) in Portland, Oregon. Dr Hersh's research focuses on the development and evaluation of information retrieval systems for biomedical practitioners and researchers.

Keywords: *text-mining, bioinformatics, natural language processing*

Aaron Michael Cohen,
Postdoctoral Fellow,
Department of Medical Informatics and Clinical Epidemiology,
School of Medicine,
Oregon Health & Science University,
3181 S.W. Sam Jackson Park Road,
Portland, OR 97239–309, USA

Tel: +1 503 494 0046
Fax: +1 503 494 4551
E-mail: cohenaa@ohsu.edu

Abstract

The volume of published biomedical research, and therefore the underlying biomedical knowledge base, is expanding at an increasing rate. Among the tools that can aid researchers in coping with this information overload are text mining and knowledge extraction. Significant progress has been made in applying text mining to named entity recognition, text classification, terminology extraction, relationship extraction and hypothesis generation. Several research groups are constructing integrated flexible text-mining systems intended for multiple uses. The major challenge of biomedical text mining over the next 5–10 years is to make these systems useful to biomedical researchers. This will require enhanced access to full text, better understanding of the feature space of biomedical literature, better methods for measuring the usefulness of systems to users, and continued cooperation with the biomedical research community to ensure that their needs are addressed.

INTRODUCTION: BACKGROUND AND PURPOSE

The volume of published biomedical research, and therefore the underlying biomedical knowledge base, is expanding at an increasing rate. While scientific information in general has been growing exponentially for several centuries,[1] the absolute numbers specific to modern medicine are very impressive. The MEDLINE 2004 database contains over 12.5 million records, and the database is currently growing at the rate of 500,000 new citations each year.[2] With such explosive growth, it is extremely challenging to keep up to date with all of the new discoveries and theories even within one's own field of biomedical research.

Biomedical research is divided into highly specialised fields and subfields, with poor communication between disciplines.[3] While this may be a necessary pre-condition for the complex and detailed research that biomedical science requires, it also tends to narrow the perspective, impeding the establishment of connections between discoveries arising from different research specialties. With the recent sequencing of the human genome, the addition of detailed genetic information to biomedical research makes the situation even more complicated, since genetics may play a role in almost all areas of health and disease and it is likely that many connections between different branches of medicine may be based on related genomic mechanisms.

The goal of biomedical research is to discover knowledge and put it to practical use in the forms of diagnosis, prevention and treatment. Clearly with the current rate of growth in published biomedical research, it becomes increasingly likely that important connections between individual elements of biomedical knowledge that could lead toward practical use are not being recognised because there is no individual in a position to make the necessary connections.

Methods must be established to aid researchers and physicians in making more efficient use of the existing research and helping them take this research to the next step along the path to practical application. While manual curation and indexing can be an aid to researchers searching for appropriate literature, a

recent study of the information content of MEDLINE records by Kostoff *et al.*[4] found a significant amount of conceptual information present only in the abstract field and missing from the MeSH terms. This is not surprising since the MEDLINE indexers and the MeSH vocabulary, while broadly based, cannot be expected to represent all of the concepts of interest for all potential users. Clearly, the full text of biomedical literature contains a wealth of information important to users that may not be completely captured by reviewers and curators.

The goal of biomedical text mining is to shift the burden of information overload from the researcher to the computer

Text mining and knowledge extraction are ways to aid researchers in coping with information overload. Text mining is differentiated from both information retrieval (IR) and text summarisation (TS) in that while IR and TS focus on the larger units of text such as documents, text mining operates at a finer level of granularity and examines the relationships between specific kinds of information contained both within and between documents. Text mining is also differentiated from full-blown natural language processing (NLP) in that NLP attempts to understand the meaning of text as a whole, while text mining and knowledge extraction concentrate on solving a specific problem in a specific domain identified *a priori* (possibly using some NLP techniques in the process). For example, text mining can aid database curators by selecting articles most likely to contain information of interest,[5,6] or potential new treatments for migraine may be determined by looking for pharmacological substances that are associated with biological processes associated with migraine.[7,8]

Recognising biological entities in text allows for further extraction of relationships and other information by identifying the key concepts of interest

The goal of biomedical text mining is therefore to allow researchers to identify needed information more efficiently, uncover relationships obscured by the sheer volume of available information, and in general shift the burden of information overload from the researcher to the computer by applying algorithmic, statistical and data management methods to the vast amount of biomedical knowledge that exists in the literature as well as the free text fields of biomedical databases.

This paper surveys the state of the art in biomedical text mining over the past 18–24 months. The next section covers current active areas of research, including the specific problems that are being addressed and the approaches used. This is followed by an examination of the current issues and future challenges of biomedical text mining.

CURRENT AREAS OF RESEARCH

While other authors have proposed categorisations based on stages of information extraction of increasing sophistication,[9] here recent work is grouped pragmatically with separate categories for each distinct type of text-mining task. This is because current work centres around several common text-mining themes.

Named entity recognition

At first glace, the task of named entity recognition (NER) appears straightforward. The goal is to identify, within a collection of text, all of the instances of a name for a specific type of thing: for example, all of the drug names within a collection of journal articles, or all of the gene names and symbols within a collection of MEDLINE abstracts. Hansich and de Bruijn and coworkers[9,10] believed that solving this problem would allow more complex text-mining tasks to be addressed. The idea is that recognising biological entities in text allows for further extraction of relationships and other information by identifying the key concepts of interest and allowing those concepts to be represented in some consistent, normalised form.

This task has been challenging for several reasons. First, there does not exist a complete dictionary for most types of biological named entities, so simple text-matching algorithms do not suffice. In addition, the same word or phrase can

refer to a different thing depending upon context (eg ferritin can be a biological substance or a laboratory test). Conversely, many biological entities have several names (eg PTEN and MMAC1 refer the same gene). Biological entities may also have multi-word names (eg carotid artery), so the problem is additionally complicated by the need to determine name boundaries and resolve overlap of candidate names.

Because of the potential utility and complexity of the problem, NER has attracted the interest of many researchers, and there is a tremendous amount of published research in this topic. With the large amount of genomic information being generated by biomedical researchers, it should not be surprising that in the genomics era, much of the work in biomedical NER has focused on recognising gene and protein names in free text.

Approaches to NER generally fall into three categories: lexicon-based, rules-based and statistically-based

The approaches generally fall into three categories: lexicon-based, rules-based and statistically based. Combined approaches also have been used. The output may be a set of tags assigning a predicted type to each word or phrase of interest, as in part-of-speech (POS) tagging,[11] or as a score designating the confidence that a word or phrase is of a given type of interest. Systems are typically measured in terms of precision (number of correct predictions divided by total number of predictions) and recall (number of correct predictions divided by number of actual named entities in the text). Precision and recall are often combined into a single measure, either using the *F*-score, defined as the harmonic mean of precision and recall (2PR/[P+R]),[12] or by reporting the balanced precision and recall, defined as the point where precision and recall are equal.

One of the most successful rules-based approaches to gene and protein NER in biomedical texts has been the AbGene system of Tanabe and Wilbur.[13] It has been used as the NER component in extracting relationships by several other researchers.[14,15] AbGene works by extending the Brill POS tagger[11,16,17] to include gene and protein names as a tag type with the system trained on 7,000 hand-tagged sentences from biomedical text. AbGene then applies manually generated post-processing rules based on lexical-statistical characteristics that help further identify the context in which gene names are used and eliminate false positives and negatives. The system achieved a precision of 85.7 per cent at a recall of 66.7 per cent.

In contrast to the tagging approach used by Tanabe and Wilbur, Chang *et al.* created the GAPSCORE system,[18] which assigns a numerical score to each word within a sentence by examining the appearance, morphology and context of the word and then applying a classifier trained on these features. Words with higher scores are more likely to be gene and protein names or symbols. After training on the Yapex corpus,[19] precision, recall and *F*-score were computed for both the exact matches and 'sloppy' matches (defined as a true positive if any part of gene name is predicted correctly), with the system performing much better with sloppy matches (precision 74 per cent, recall 81 per cent, *F*-measure 77 per cent), than with exact matches (precision 59 per cent, recall 50 per cent, *F*-measure 54 per cent).

A number of other groups have worked in this area. Hanisch *et al.* used a large dictionary of gene and protein names and semantically classified words that tend to appear in context with protein names, reporting a specificity of 95 per cent and sensitivity of 90 per cent.[10] Zhou *et al.* trained a hidden Markov model (HMM) on a set of features based on word formation (ie capitalisation), morphology (ie prefix and suffix), POS, semantic triggers (head nouns and verbs) and intra-document name aliases.[20] They reported an overall precision of 66.5 per cent at a recall of 66.6 per cent on the GENIA corpus.[21] Other gene and protein NER systems include those by Narayanaswamy *et al.*,[22] Settles [23] and Mika and Rost.[24]

Chen and Friedman have adapted the MEDLEE system to recognise phrases that correspond to phenotype information within biomedical text.[25] This system uses natural language techniques to identify phenotypic phrases present in journal article abstracts, and recognises phrases containing words separated in the text. This area of biological NER is much less well studied than recognising gene, protein or chemical names, and therefore a smaller knowledge base of phenotype-associated terms is available. Nevertheless, the investigators were able to automatically import thousands of UMLS terms associated with semantic categories such as cellular body functions and cellular dysfunction, as well as several hundred terms from the Mammalian Ontology. A few hundred other terms were added manually. In a feasibility study of 300 documents, the system achieved a precision of 64.0 per cent with a recall of 77.1 per cent. While, as expected for a new area of study, this performance is lower than that of the gene and protein NER systems, these results were found to be about the same as that of the individual experts used to create the study's gold standard.

Overall, the performance of state-of-the-art gene and protein NER systems, achieves F-scores between 75 and 85 percent

Overall, the performance of state-of-the-art gene and protein NER systems achieves *F*-scores between 75 and 85 per cent. This number is consistent with that found by Hirschman *et al*. in 2002,[12] and the results of Task 1A for the 2004 BioCreative workshop.[15] While peak performance does not appear to have increased over the past few years, investigators are obtaining consistent results using a variety of approaches on different data sets.

To address this performance plateau and to decrease the computational burden contribution of NER to text mining, Tanabe and Wilbur have used AbGene to generate a large and high-quality gene and protein lexicon of names found in biomedical text.[26] The application of AbGene to the MEDLINE database has resulted in an initial collection of over two million putative gene and protein names. This list was purified by applying thematic analysis to the names, and then using inductive logic programming to learn rules for differentiating gene names from non-gene names within a theme. Finally, a simple false-positive filter was applied that removed obviously incorrect names such as those containing 'http' or ending in 'tion'. Their approach yielded a final set of 1,145,913 gene names. Assessment of a random sample determined the precision to be approximately 82 per cent. Comparison with a gold standard gave an estimated recall of 61 per cent for exact matches and 88 per cent for partial matches.

The quality of this lexicon is about equivalent to the performance of the NER systems, and the large size of the lexicon is a definite advantage. The lexicon could be used with simple or fuzzy matching to efficiently identify gene and protein names as a first step in future text-mining systems. However, the list has been generated with a snapshot of MEDLINE, and given the pace of genomic research, will soon be out of date if it is not updated. Also, the list was built from MEDLINE and not from full text articles, so it is possible that a significant number of gene and protein names exist in the literature and are not found in MEDLINE.

It is a subject of current debate how well NER must perform in order to be useful for text mining.[9,12] If one assumes that relationship extraction requires identification of three biomedical terms (two entities and one relationship), the performance of relationship extraction should be approximately equal to the cube of the performance of NER. This *independence assumption* appears to be true for news article extraction. Systems performing named entity extraction on news stories typically perform at an *F*-score over 90 per cent, and the *F*-score for new relations is about 75 per cent.

For many biomedical applications, the *F*-score performance rates of relationship mining has often been found to be approximately equal to that of biological

The independence assumption does not seem to hold for biological relations

NER, rather than the 60 per cent expected by the independence assumption.[27–32] Therefore, the assumption does not seem to hold for biological relations. It may be easier to extract concepts in combination with the relationship between them owing to the increased local context that relationships provide. While some form of NER is useful in most text-mining tasks, the performance level of biological NER is not necessarily rate limiting for other biological text-mining tasks. Nevertheless, we have not reached the point of having standard methods of NER or updated lexicons for biomedical text mining (whatever the asymptotic performance level), so work must continue in this area.

Text classification

Text classification attempts to automatically determine whether a document or part of a document has particular characteristics of interest, usually based on whether the document discusses a given topic or contains a certain type of information. Typically the information of interest is not specified explicitly by the users and, instead, they provide a set of documents that have been found to contain the characteristics of interest (the positive training set), and another set that does not (the negative training set). Text classification systems must automatically extract the features that help determine positives from negatives and apply those features to candidate documents using some kind of decision-making process.

Accurate text classification systems can be especially valuable to database curators

Accurate text classification systems can be especially valuable to database curators, who may have to review many documents to find a few that contain the kind of information they are collecting in their database. Because more biomedical information is being created in text form then ever before, and because there are more ongoing database curation efforts to organise this information into coded databases than before, there is a strong need to find useful ways to apply text classification methods to biomedical text.

Yeh *et al.* ran a text-mining competition as part of the *Knowledge Discovery in Databases* (KDD) Challenge Cup 2002.[6] The task was a curation problem to evaluate papers from the FlyBase data set and determine whether the paper should be curated based on the presence of experimental evidence of *Drosophila* gene products. The best-performing entry used a set of manually constructed rules based on POS tagging, a lexicon and semantic constraints determined by examining the training documents.[33] The system focused on figure captions, which were found to be useful. An *F*-score of 78 per cent was achieved on determining whether to curate a paper based on the presence of experimental evidence. Another effective approach looked for manually chosen 'keywords' and computed the distance between keywords and gene names.[34] Two other well-performing systems used regular expressions to find patterns of words and then used a support vector machine (SVM) to classify the papers.[35]

In related work, Donaldson *et al.* used an SVM trained on the words in MEDLINE abstracts to distinguish abstracts containing information on protein–protein interactions, prior to curating this information into their BIND database.[36] They used the 'bag-of-words' approach with an SVM classifier. A small evaluation with 100 abstracts found a precision of 96 per cent with a recall of 84 per cent. They estimated that the classification system would reduce the number of abstracts that the curators needed to read by about two-thirds.

Another investigation in this area used a Probabilistic Latent Categoriser (PLC) with Kullback–Leibler (KL) divergence to re-rank documents returned by PubMed searching for the purposes of curating information into the Swiss-Prot database.[37] Evaluation showed a 25–45 per cent precision improvement, with a balanced precision and recall point of about 70 per cent, compared with about 40 per cent for the basic PubMed ranking. Liu *et al.* performed a unique application

of text classification on figure captions. In a pilot study, they classified the text in figure legends in order to find figures containing representations of protein interactions and signalling events.[38]

Applying research in text classification to the work processes of actual biomedical curators and annotators is just beginning. The Text Retrieval Conference (TREC) 2004 Genomics Track has a classification problem as one of its tasks.[39] The task is meant to mimic the process that the human annotators in the Mouse Genome Informatics (MGI) system go through in order to find documents that contain experimental evidence about genes that they are annotating using Gene Ontology (GO) codes. A full text collection in SGML format has been assembled, realistically reflecting the articles that MGI annotators currently read. In additional, the utility measure used to evaluate performance of the task aims to reflect the priorities of the MGI annotators. Because of the potential to improve annotator productivity, work on improving biomedical text classification to meet the needs of curators and other users must continue for the foreseeable future.

Because of the potential to improve annotator productivity, work on improving biomedical text classification must continue for the foreseeable future

Synonym and abbreviation extraction

Paralleling the growth of the increase in biomedical literature is the growth in biomedical terminology. Because many biomedical entities have multiple names and abbreviations, it would be advantageous to have an automated means to collect these synonyms and abbreviations to aid users doing literature searches. Furthermore, other text-mining tasks could be done more efficiently if all of the synonyms and abbreviations for an entity could be mapped to a single term representing the concept. Most of the work in this type of extraction has focused on uncovering gene name synonyms and biomedical term abbreviations.

Several investigators have used gene name synonym lists created from online databases as a basis for further text mining.[36,40,41] However, these gene databases focus on official names and alternates, and are incomplete with respect to the gene names actually found in the literature.[42,43] In order to create gene and protein name synonym lists representative of the names used in the literature, Yu and Agichtein[44] and Cohen[45] have investigated automatic means of extracting gene name synonyms from biomedical free text. Yu and Agichtein applied a combination of four algorithms to full text journal articles. Their system combined the AbGene gene NER system, with statistical, SVM classifier-based, automatic pattern-based and manual rules algorithms. The combined system produced a recall of about 80 per cent with a precision of about 9 per cent, giving an overall *F*-measure of about 30 per cent. Cohen applied an automatic pattern extraction method to MEDLINE abstracts and a numeric analysis metric on the resulting name co-occurrence network to select the best synonym extraction patterns. While no sophisticated gene NER was used, evaluation showed a precision of 23 per cent, a recall of 21 per cent and an *F*-score of 22 per cent. The system was also notable for inferring synonyms based on the logical relationship between synonyms found explicitly in the text, increasing recall by about 10 per cent over the same system without inference.

Other investigators have applied text-mining methods to extracting lists of biomedical abbreviations and their fully specified forms. These methods rely on the proximity of full forms and their abbreviations, and the fact that either the full form or the abbreviation is often enclosed in parentheses. The problem is often reduced to finding the best alignment of the characters in the abbreviation to those in the full form. A variety of alignment and scoring methods have been applied to this basic approach. Liu and Friedman used a large

collection of MEDLINE abstracts to determine abbreviations and phrases that were statistically significantly co-located.[32] They reported a precision of 96.3 per cent with a recall of 88.5 per cent. Yu *et al.*[46] and Schwartz and Hearst[28] applied manually created set of pattern-matching rules to identify abbreviations and their full form. Yu *et al.* achieved a precision of 95 per cent with 70 per cent recall, while Schwartz and Hearst achieved a precision of 96 per cent at 82 per cent recall for a set of 1,000 MEDLINE abstracts mentioning yeast. Chang *et al.* trained a logistic regression model with abbreviation specific features and used it to score candidate full forms,[47] achieving a precision of 80 per cent with 83 per cent recall on the Medstract corpus.[48]

The automatic extraction of biomedical abbreviations and their corresponding definition as used within an individual journal article is close to being a solved problem. Research systems uniformly produce high precision and recall. The next step is to integrate these automated extraction capabilities into user systems. For example, an online dictionary of medical abbreviations could be integrated into PubMed to augment search queries. The more general problem of resolving common domain abbreviations undefined in a given journal article is a much more difficult problem dependent on expert knowledge of the specific field and additional, possibly subtle, context from the surrounding text.

It is thought that, grouping genes by functional relationships could aid gene expression analysis and database annotation

Gene and protein name synonym extraction has proven to be a more challenging problem. While an automatically updated synonym list would be of great value in augmenting literature searching and text mining, the precision of automatic extraction systems is low enough to introduce an unacceptable level of noise. However, work is being undertaken to standardise the use of official gene and protein names and symbols,[43] so this problem may lessen in the future. On the other hand, there will still be a large legacy of literature that uses non-official names.

Relationship extraction

The goal of relationship extraction is to detect occurrences of a prespecified type of relationship between a pair of entities of given types. While the type of the entities is usually very specific (eg genes, proteins or drugs), the type of relationship may be very general (eg any biochemical association) or very specific (eg a regulatory relationship). Several approaches to extracting relations of interest have been reported in the literature and are applicable to this work. Manually generated template-based methods use patterns (usually in the form of regular expressions) generated by domain experts to extract concepts connected by a specific relation from text.[14] Automatic template methods create similar templates automatically by generalising patterns from text surrounding concept pairs known to have the relationship of interest.[44,45] Statistical methods identify relationships by looking for concepts that are found with each other more often than would be predicted by chance.[7] Finally, NLP-based methods perform a substantial amount of sentence parsing to decompose the text into a structure from which relationships can be readily extracted.[31]

In the current genomic era, most investigation of this type has centred around relationships between genes and proteins. It is thought that grouping genes by functional relationships could aid gene expression analysis and database annotation.[49] Several researchers have investigated the extraction of general relationships between genes.

Genes can be grouped or clustered based on how strongly they share words in text containing their names. Raychaudhuri *et al.* used a measure of neighbour divergence to measure the 'functional coherence' of a group of genes.[49] They obtained 79 per cent sensitivity at 100 per cent specificity for distinguishing 19 true gene groups from

1,900 randomly assembled groups of yeast genes. They later extended their work to include mouse, fly, worm and yeast genes and obtained functional gene groups with 96, 92, 82 and 45 per cent sensitivity at 99.9 per cent specificity.[50] Glenisson *et al.* similarly investigated text-based gene clustering using a vector space approach and the *k*-medoids algorithm with a cosine similarity metric.[51] Wren and Garner identified related genes by analysing the cohesiveness and specificity of the graph structure created by the gene–gene co-occurrences in MEDLINE records.[27] They obtained similar results to Raychaudhuri *et al.* of about 97 per cent specificity at 85 per cent sensitivity.

Currently, the precision and recall obtained for relationship extraction is dependent on the type of relationship to be extracted and literature corpus to be processed

Other research has concentrated extracting specific kinds of relationships between genes, protein, or other biological entities. Gaizauskas *et al.*'s Protein Active Site Template Acquisition system (PASTA) uses type and POS tagging along with manually created templates and lexicons assembled from biological databases to extract relationships between amino acid residues and their function within a protein.[30] Balanced recall and precision was approximately 82 per cent using a manually annotated corpus of MEDLINE abstracts as a gold standard. Albert *et al.* used dictionaries of protein and interaction terms to identify tri-occurrences of two proteins and one interaction within a sentence.[41] Applying this approach to the full MEDLINE database looking for interactions between proteins and nuclear receptors they found 3,308 positive interactions, giving a precision of 22 per cent. McDonald *et al.* combined a hybrid syntactic/semantic grammar in a single parsing process to extract a variety of gene pathway relationships.[29] Evaluation using 100 abstracts manually reviewed by a biologist showed 61 per cent precision at 35 per cent recall.

Extracting relationships between genes or proteins and GO codes is a task with immediate practical potential that has received much attention lately. The MeKE system of Chiang and Yu used GO codes as a lexicon of function names, combining it with a lexicon of gene and gene product names from LocusLink, and used a sentence alignment system to determine patterns associated with statements about gene function. They then used the patterns with a Naive Bayes classifier to extract sentences containing information about gene product function.[52]

Raychaudhuri *et al.* assigned GO codes by training text classifiers to associate GO codes with abstracts, and then assigning to a gene the strongest maximum entropy associated GO code from the abstracts in which that gene appeared. Evaluation using a subset of yeast genes and GO codes showed that the strongest predicted GO code was accurate about 72 per cent of the time.[53] Pan *et al.*'s Dragon TF association miner system used linear discriminate analysis on terms and neural networks to create models that recognised abstracts that contained information relating transcription factors (TFs) with GO codes and diseases. Balanced sensitivity and specificity was about 80 per cent.[54]

Task 2 of the BioCreative 2004 workshop similarly focused on extracting relevant GO codes for genes from free text.[15,55] The task was to identify text that contained evidence for GO code assignment and to predict the correct GO code that should be assigned by that text. The task was also notable in that full text journal articles were used and the evaluation was carried out by MGI curators, who rated how useful the system output was for annotation. In general, this was a difficult task, and was rated very hard by the annotators. System precision ranged between 2 and 80 per cent, with the average about 30 per cent. Recall was not evaluated. Part of the difficulty was that the systems had to get the text, the gene and the GO code simultaneously all correct, perhaps an unnecessarily high standard.

A number of other investigators have applied text mining to extract novel, interesting relationships. Eskin and

Agichtein combined text and sequence mining with an SVM combined text and genome sequence kernel to predict protein subcellular localisation.[56] Performance ranged from a precision of 87 per cent with 71 per cent recall for proteins located in the cytoplasm to a precision of 44 per cent with 21 per cent recall for proteins located in the peroxisomes. Srinivasan and Wedemeyer have studied the relationship between a disease's incidence and the countries in which it is studied.[57] Kostoff used simple MEDLINE querying to compute organ cancer asymmetries and found very similar results to numbers in the National Cancer Institutes SEER database.[58] Xu *et al.* adapted MEDLEE to transform text into coded data from pathology reports to facilitate a breast cancer study.[59]

Extraction of very general, non-specific relationships appears to be straightforward

It is clear from the foregoing work that some types of relationships are simpler to extract than others. Very general, non-specific relationships (eg gene groups) seem to be fairly straightforward, while very specific relationships that have to be substantiated by the precise supporting text (eg GO code assignment) remain challenging. Since the value of identifying very specific relationships with accompanying supporting text is high, this work must receive continued attention.

Hypothesis generation

While relationship extraction focuses on the extraction of relationships between entities explicitly found in the text, hypothesis generation attempts to uncover relationships that are not present in the text but instead are inferred by the presence of other more explicit relationships. The goal is to uncover previously unrecognised relationships worthy of further investigation.

Current work in hypothesis generation makes use of 'complementary structures in disjoint literatures'

Practically all of the work in hypothesis generation makes use of an idea originated by Swanson in the 1980s called the 'complementary structures in disjoint literatures' (CSD).[60] Swanson realised that large databases of scientific literature could allow discoveries to be made by connecting concepts using logical inference. He proposed a simple 'A influences B, and B influences C, therefore A may influence C' model for detecting instances of CSD that is commonly referred to as *Swanson's ABC model*.[3,61] In several published papers in the 1980s and early 1990s, Swanson gave examples of discovering new hypotheses by manually connecting concepts between journal articles. In 1986, he found a connection implying patient benefit between fish oil and Raynaud's syndrome, two years before clinical trials established that the benefit was real.[8,62] In another article, he traced 11 indirect connections between migraine and magnesium using summarisations of published articles[60] that were later experimentally verified.[63,64]

While Swanson applied his model manually, several investigators have tried to automate the process. Automated hypothesis generation systems may generate many potential hypotheses, and therefore some method of evaluating these systems is necessary. One way these evaluations have been done is by attempting to recreate Swanson's discoveries. Gordon and Lindsay were probably the first to use this approach,[65] followed a few years later by Weeber *et al.*[3] More recently, Srinivasan has used this approach to demonstrate the feasibility of her approach based on MeSH terms and UMLS semantic types.[66]

Another way that hypothesis discovery systems are evaluated is by manually reviewing the literature supporting the extracted hypothesis for scientific plausibility and relevance. This is a natural next step after reproducing Swanson's discoveries. Using their 'literature-based scientific discovery tool' that examined term co-occurrences in MEDLINE titles and abstracts, Weeber *et al.* found potential new uses for thalidomide.[61] Srinivasin *et al.* continued to refine their system, discovering implicit evidence of a therapeutic effect of *Curcuma longa* (turmeric), on retinal diseases, Crohn's disease and spinal cord injuries.[67,68]

The combined increase in scientific

literature and genome expression data may leave many scientists with the uneasy feeling that important discoveries are buried under the information explosion, such that computerised tools to help them sort through the vast amount of available information.[12,68] While hypothesis discovery systems are not yet a standard tool of biologists, one day they may be. Continued work is needed to enhance these systems to handle the vast amounts of different types of data that scientists currently must explore manually. Additionally, better methods are needed to evaluate and compare the results of these systems so that improvement can be documented and clear choices can be made.

Hypothesis discovery systems are not yet a standard tool of biologists

Integration frameworks

Several research groups are developing integrated text-mining frameworks intended to be able to address a variety of user needs. The MedScan system of Novichkova *et al.* combines lexicons with syntactic and semantic templates into a general-purpose text-mining system to extract relationships between biomedical entities.[69] Glenisson *et al.* have developed TXTGate which performs gene-based text profiling and clustering using the information contained on multiple on-line biological databases.[70] Becker *et al.* created PubMatrix, a tool that displays two-dimensional comparisons of gene names and functional terms based on combining the results of multiple queries to PubMed.[71] The BioRAT system of Corney *et al.* is another template-based system that combines a template design tool with a web spider that locates and retrieves full text journal articles.[72] The Textpresso system of Müller *et al.* uses a specially created ontology to flexibly combine keyword and concept co-occurrence searching of *Caenorhabditis elegans* (a small nematode worm) literature at the sentence level.[73] Other generalised text-mining frameworks have been reported by Nenadic *et al.*[74] and Chiang *et al.*[40]

Greater access to full text collections is essential to progress in biomedical text-mining research

All these systems are still in the research and development phase. At this point evaluations tend to be brief and these systems have not been subjected to thorough user evaluations. It remains to be seen whether these systems will address the needs of the biomedical research community, but it is clear that they are a step towards addressing the needs of biomedical researchers beyond those served by search engines such as PubMed and Google.

CHALLENGES, FUTURE DIRECTIONS AND CONCLUSIONS

From all of the foregoing, it is clear that biomedical text mining has great potential. However, that potential is yet unrealised. Text-mining tools are not part of the standard arsenal of the biomedical researcher in the way that search engines and sequence alignment tools are. The major challenge for the next 5–10 years of text-mining work is the creation of text-mining tools to provide a clear benefit to these researchers, allowing them to be more productive given increasing challenges due to information growth. The focus must be more on helping biomedical researchers to solve real-world problems that are inhibiting the pace of research and less on evaluations based on system output independent of meeting user needs. Advances on several fronts are necessary for this to become a reality.

First, there must be greater access to full text and test collections that use it. Much of the scientific information contained in journal articles is not present or mentioned in abstracts or MeSH terms.[14] Text-mining research has recently been moving away from abstracts and titles towards full text, but access to full text is still often limited by copyright restrictions,[39,55] preventing work from being done or reproduced. The research community must work with publishers to make a wider range of content available to text-mining applications.

Next, more work is necessary to

determine what features and types of features are useful in addressing particular text-mining tasks. The feature space available to text mining is large, and includes a huge array of feature types including (but not limited to) words, concepts, headings, formatting, authors, references and links. The bag-of-words approach (both stemmed and unstemmed) has been popular for quite a long time, largely because it is easily applied to text from a large variety of sources. However, this approach ignores document positional and sectional information and may not result in the most discriminating feature set from fully marked-up text that provides this information. Using concepts as a basic unit instead of words has been shown to be practical[75] and useful.[76,77] Including other contexts, such as the section in which the text occurs, has also been shown to help.[44,78] Full text with XML mark-up may provide many more possible features and feature types than plain text. The potential feature space of XML full text has only begun to be explored. With such a wide range of possible features and feature types, additional analytical methods are needed to determine the optimal feature set for a particular application.

Better ways are needed to assess the real-world value of text-mining systems to their intended users

Text-mining researchers also need to better understand what measures can be used to assess value to actual users, and how to tailor their algorithms to meet their needs. It is well known in information retrieval that increases in precision or recall do not necessarily correlate with user success in the searching task.[79] As such, simply optimising these metrics may not result in systems that meet users' needs. The triage task of the TREC 2004 Genomics Track begins to address this with an evaluation measure based on estimating the utility of MGI's current triage process.[39] Other researchers have addressed this as well.[36] In order to design systems that deliver value to biomedical researchers, much more work is needed in the area of creating evaluation metrics and methods that measure the real-world value of text-mining systems. This is an especially challenging problem for hypothesis generation systems, eg how does one measure the value of an untested set of hypotheses? Nevertheless, robust evaluation measures are necessary to compare hypothesis generation systems, determine the best approaches and improve the state of the art of these systems. Verifying that these systems produce suggestions that biomedical researchers are motivated to experimentally test is especially important.

Finally, the approach of shared challenge tasks with consistent evaluation based on biomedical domain expertise must continue. More progress must be made toward choosing tasks and evaluating results based on real-world needs. Recent examples of this type of cooperation include the BioCreative 2004 workshop, and the TREC Genomics Track, both of which used assessments made by biological database curators in their normal workflow processes as the gold standard.

Clearly, the main theme for future progress is interdisciplinary coordination and cooperation. Text-mining researchers must work with each other, publishers and biomedical researchers to begin to meet user needs with systems that produce consistent, measurable and verifiable results. This is an exciting time in biomedical text mining, full of promise. Researchers must lead the coordination effort to realise the full scientific potential of biomedical text mining.

Annotated bibliography

⋆ Papers of particular interest published within the period of this review.

⋆⋆ Papers of extreme interest published within the period of this review.

1. ⋆Yeh, A. S., Hirschman, L. and Morgan, A. A. (2003), 'Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup', *Bioinformatics*, Vol. 19 Suppl. 1, pp. i331–339.
2. ⋆⋆Tanabe, L. and Wilbur, W. J. (2002), 'Tagging gene and protein names in biomedical text', *Bioinformatics*, Vol. 18(8), pp. 1124–1132.
3. ⋆Chang, J. T., Schutze, H. and Altman, R. B.

(2004), 'GAPSCORE: Finding gene and protein names one word at a time', *Bioinformatics*, Vol. 20(2), pp. 216–225.

4. **Hirschman, L., Morgan, A. A. and Yeh, A. S. (2002), 'Rutabaga by any other name: Extracting biological names', *J. Biomed. Inform.*, Vol. 35(4), pp. 247–259.

5. **Tanabe, L. and Wilbur, W. J. (2004), 'Generation of a large gene/protein lexicon by morphological pattern analysis', *J. Bioinform. Comput. Biol.*, Vol. 1(4), pp. 611–626.

6. *Schwartz, A. S. and Hearst, M. A. (2003), 'A simple algorithm for identifying abbreviation definitions in biomedical text', in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 451–462.

7. *Liu, H. and Friedman, C. (2003), 'Mining terminological knowledge in large biomedical corpora', in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 415–426.

8. *Gaizauskas, R., Demetriou, G., Artymiuk, P. J. and Willett, P. (2003), 'Protein structures and information extraction from biological texts: The PASTA system', *Bioinformatics*, Vol. 19(1), pp. 135–143.

9. *Friedman, C., Kra, P., Yu, H. *et al.* (2001), 'GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles', *Bioinformatics*, Vol. 17, Suppl. 1, pp. S74–82.

10. *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 Bioinformatics*, Vol. 4(1), p. 11.

11. *Dobrokhotov, P. B., Goutte, C., Veuthey, A. and Gaussier, E. (2003), 'Combining NLP and probabilistic categorisation of document and term selection for Swiss-Prot medical annotation', *Bioinformatics*, Vol. 19, Suppl. 1, pp. i91–94.

12. *Oregon Health & Science University (2004), 'TREC Genomics Track Protocol, 2004' (URL: http://medir.ohsu.edu/~genomics/2004protocol.html, accessed 27th August, 2004).

13. *The Human Genome Organisation (2003), 'HUGO Gene Nomenclature Committee, 2003' (URL: http://www.gene.ucl.ac.uk/nomenclature/, accessed 29th September, 2003).

14. *Yu, H. and Agichtein, E. (2003), 'Extracting synonymous gene and protein terms from biological literature', *Bioinformatics*, Vol. 19, Suppl. 1, pp. i340–349.

15. *Chang, J. T., Schutze, H. and Altman, R. B. (2002), 'Creating an online dictionary of abbreviations from MEDLINE', *J. Amer. Med. Inform. Assoc.*, Vol. 9(6), pp. 612–620.

16. *Chiang, J. H. and Yu, H. C. (2003), 'MeKE: Discovering the functions of gene products from biomedical literature via sentence alignment', *Bioinformatics*, Vol. 19(11), pp. 1417–1422.

17. **Eskin, E. and Agichtein, E. (2004), 'Combining text mining and sequence analysis to discover protein functional regions', in 'Proceedings of the 9th Pacific Symposium on Biocomputing', 6th–10th January, Hawaii, pp. 288–299.

18. *Swanson, D. R. (1991), 'Complementary structures in disjoint science literatures', in 'Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM Press, Chicago, IL, pp. 280–289.

19. *Weeber, M., Vos, R., Klein, H. *et al.* (2003), ,Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide', *J. Amer. Med. Inform. Assoc.*, Vol. 10(3), pp. 252–259.

20. *Srinivasan, P. (2004), 'Text mining: Generating hypothesis from MEDLINE', *J. Amer. Soc Inf. Sci. Technol.*, Vol. 55, pp. 396–413.

21. *Srinivasan, P., Libbus, B. and Sehgal, A. K. (2004), 'Mining MEDLINE: Postulating a beneficial role for curcumin longa in retinal diseases', in 'BioLINK 2004: Linking Biological Literature, Ontologies, and Databases', Boston, MA, Association for Computational Linguistics, p. 33–40 (URL: http://www.cs.brandeis.edu/~jamesp/biolink2004/).

22. *Corney, D. P., Buxton, B. F., Langdon, W. B. and Jones, D. T.(2004), 'BioRAT: Extracting biological information from full-length papers', *Bioinformatics* (in press).

23. *Aronson, A. R. (2001), 'Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program', in 'Proceedings of the AMIA Symposium', 3rd–7th November, Washington, DC, pp. 17–21.

24. *Müller, H., 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).

References

1. Hersh, W. R. (2003), 'Information Retrieval: A Health and Biomedical Perspective', 2nd edn, Springer, New York.

2. Mitchell, J. A., Aronson, A. R., Mork, J. G. *et al.* (2003), 'Gene indexing: Characterization and analysis of NLM's GeneRIFs', in 'Proceedings of the AMIA Symposium', 8th–12th November, Washington, DC, pp. 460–464.

3. Weeber, M., Klein, H., Aronson, A. R. *et al.*

(2000), 'Text-based discovery in biomedicine: The architecture of the DAD-system', in 'Proc. AMIA Symposium', 4th–8th November, Los Angeles, CA, pp. 903–907.

4. Kostoff, R. N., Block, J. A., Stump, J. A. and Pfeil, K. M. (2004), 'Information content in Medline record fields', *Int. J. Med. Inf.*, Vol. 73(6), pp. 515–527.
5. Hersh, W., Bhupatiraju, R. T., Oregon Health & Science University (2004), 'TREC Genomics Track Overview, 2003' (URL: http://medir.ohsu.edu/~genomics/overview.pdf, accessed 13th September, 2004).
6. Yeh, A. S., Hirschman, L. and Morgan, A. A. (2003), 'Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup', *Bioinformatics*, Vol. 19, Suppl. 1, pp. i331–339.
7. Lindsay, R. K. and Gordon, M. D. (1999), 'Literature-based discovery by lexical statistics', *J. Amer. Soc. Information Sci.*, Vol. 50(7), pp. 574–587.
8. Swanson, D. R. (1990), 'Medical literature as a potential source of new knowledge', *Bull. Med. Libr. Assoc.*, Vol. 78(1), pp. 29–37.
9. de Bruijn, B. and Martin, J. (2002), 'Getting to the (c)ore of knowledge: Mining biomedical literature', *Int. J. Med. Inf.*, Vol. 67(1–3), pp. 7–18.
10. 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', 3rd–7th January, Hawaii, pp. 403–414.
11. Brill, E. (1995), 'Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging', *Comput. Linguistics*, Vol. 21(4), pp. 543–565.
12. Hirschman, L., Morgan, A. A. and Yeh, A. S. (2002), 'Rutabaga by any other name: Extracting biological names', *J. Biomed. Inform.*, Vol. 35(4), pp. 247–259.
13. Tanabe, L. and Wilbur, W. J. (2002), 'Tagging gene and protein names in biomedical text', *Bioinformatics*, Vol. 18(8), pp. 1124–1132.
14. Yu, H., Hatzivassiloglou, V., Friedman, C. *et al.* (2002), 'Automatic extraction of gene and protein synonyms from MEDLINE and journal articles', in 'Proceedings of the AMIA Symposium', 9th–13th November, San Antonio, TX, pp. 919–923.
15. Blaschke, C. (2004), 'Biocreative: Critical Assessment for Information Extraction in Biology', Granada, Spain (URL: http://www.pdg.cnb.uam.es/Biolink/Workshop_BioCreative_04/handout/index.html).
16. Brill, E. (2003), 'Processing natural language without natural language processing', in 'Computational Linguistics and Intelligent Text Processing, Proceedings', Lecture Notes in Computer Science, Vol. 2588, Gelbukh, A. F., Ed, Springer, pp. 360–369.
17. Brill, E. and Mooney, R. J. (1997), 'An overview of empirical natural language processing', *AI Magazine*, Vol. 18(4), pp. 13–24.
18. Chang, J. T., Schutze, H. and Altman, R. B. (2004), 'GAPSCORE: Finding gene and protein names one word at a time', *Bioinformatics*, Vol. 20(2), pp. 216–225.
19. Franzen, K., Eriksson, G., Olsson, F. *et al.* (2002), 'Protein names and how to find them', *Int. J. Med. Inf.*, Vol. 67(1–3), pp. 49–61.
20. Zhou, G., Zhang, J., Su, J. *et al.* (2004), 'Recognizing names in biomedical texts: A machine learning approach', *Bioinformatics*, Vol. 20(7), pp. 1178–1190.
21. Kim, J. D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003), 'GENIA corpus – a semantically annotated corpus for bio-textmining', *Bioinformatics*, Vol. 19, Suppl. 1, pp. i180–182.
22. Narayanaswamy, M., Ravikumar, K. E. and Vijay-Shanker, K. (2003), 'A biological named entity recognizer', in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 427–38.
23. Settles, B. (2004), 'Biomedical named entity recognition using conditional random fields and rich feature sets', in 'Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)', Geneva, Switzerland.
24. Mika, S. and Rost, B. (2004), 'Protein names precisely peeled off free text', *Bioinformatics*, Vol. 20, Suppl. 1, pp. I241–247.
25. Chen, L. and Friedman, C. (2004), 'Extracting phenotypic information from the literature via natural language processing', in 'Proceedings of the 11th World Congress on Medical Informatics', IMIA, San Francisco, CA, pp. 758–762.
26. Tanabe, L. and Wilbur, W. J. (2004), 'Generation of a large gene/protein lexicon by morphological pattern analysis', *J. Bioinform. Comput. Biol.*, Vol. 1(4), pp. 611–626.
27. Wren, J. D. and Garner, H. R. (2004), 'Shared relationship analysis: Ranking set cohesion and commonalities within a literature-derived relationship network', *Bioinformatics*, Vol. 20(2), pp. 191–198.
28. Schwartz, A. S. and Hearst, M. A. (2003), 'A simple algorithm for identifying abbreviation definitions in biomedical text', in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 451–462.
29. McDonald, D. M., Chen, H., Su, H. and Marshall, B. B. (2004), 'Extracting gene

pathway relations using a hybrid grammar: The Arizona relation parser', *Bioinformatics* (in press).

30. Gaizauskas, R., Demetriou, G., Artymiuk, P. J. and Willett, P. (2003), 'Protein structures and information extraction from biological texts: The PASTA system', *Bioinformatics*, Vol. 19(1), pp. 135–143.
31. Friedman, C., Kra, P., Yu, H. *et al.* (2001), 'GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles', *Bioinformatics*, Vol. 17, Suppl. 1, pp. S74–82.
32. Liu, H. and Friedman, C. (2003), 'Mining terminological knowledge in large biomedical corpora', in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 415–426.
33. Regev, Y., Finkelstein-Landau, M. and Feldman, R. (2002), 'Rule-based extraction of experimental evidence in the biomedical domain: The KDD Cup 2002 (task 1)', *ACM SIGKDD Explorations Newsletter*, Vol. 4(2), pp. 90–92.
34. Shi, M., Edwin, D. S., Menon, R. *et al.* (2002), 'A machine learning approach for the curation of biomedical literature-KDD Cup 2002 (task 1)', *ACM SIGKDD Explorations Newsletter*, Vol. 4(2), pp. 93–94.
35. Ghanem, M. M., Guo, Y., Lodhi, H. and Zhang, Y. (2003), 'Automatic scientific text classification using local patterns: KDD Cup 2002 (task 1)', *ACM SIGKDD Explorations Newsletter*, Vol. 4(2), pp. 95–96.
36. 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 Bioinformatics*, Vol. 4(1), p. 11.
37. Dobrokhotov, P. B., Goutte, C., Veuthey, A. and Gaussier, E. (2003), 'Combining NLP and probabilistic categorisation of document and term selection for Swiss-Prot medical annotation', *Bioinformatics*, Vol. 19, Suppl. 1, pp. i91–94.
38. Liu, F., Jenssen, T. K., Nygaard, V. *et al.* (2004), 'FigSearch: A figure legend indexing and classification system', *Bioinformatics*, Vol. 20, pp. 2880–2882.
39. Hersh, W.R. and Oregon Health & Science University (2004), 'TREC Genomics Track Protocol' (URL: http://medir.ohsu.edu/~genomics/2004protocol.html, accessed 27th August, 2004).
40. Chiang, J. H., Yu, H. C. and Hsu, H. J. (2004), 'GIS: A biomedical text-mining system for gene information discovery', *Bioinformatics*, Vol. 20(1), pp. 120–121.
41. Albert, S., Gaudan, S., Knigge, H. *et al.* (2003), 'Computer-assisted generation of a protein-interaction database for nuclear receptors', *Mol. Endocrinol.*, Vol. 17(8), pp. 1555–1567.
42. Editorial (2003), 'HUGO – a UN for the human genome', *Nat. Genet.*, Vol. 34(2), pp. 115–116.
43. The Human Genome Organisation, HUGO Gene Nomenclature Committee, 2003 (URL: http://www.gene.ucl.ac.uk/nomenclature/, accessed 29th September, 2003).
44. Yu, H. and Agichtein, E. (2003), 'Extracting synonymous gene and protein terms from biological literature', *Bioinformatics*, Vol. 19, Suppl. 1, pp. i340–349.
45. Cohen, A. M. (2004), 'Using symbolic network logical analysis as a knowledge extraction method on MEDLINE abstracts', BMC Bioinformatics 2005 (in press).
46. Yu, H., Hripcsak, G. and Friedman, C. (2002), 'Mapping abbreviations to full forms in biomedical articles', *J. Amer. Med. Inform. Assoc.*, Vol. 9(3), pp. 262–272.
47. Chang, J. T., Schutze, H. and Altman, R. B. (2002), 'Creating an online dictionary of abbreviations from MEDLINE', *J. Amer. Med. Inform. Assoc.*, Vol. 9(6), pp. 612–620.
48. Brandeis University. Medstract Project – Initial Annotion Corpora (2001) (URL: http://scylla.cs.brandeis.edu/gold-standards.html, accessed 24th June, 2003).
49. Raychaudhuri, S., Schutze, H. and Altman, R. B. (2002), 'Using text analysis to identify functionally coherent gene groups', *Genome Res.*, Vol. 12(10), pp. 1582–1590.
50. Raychaudhuri, S. and Altman, R. B. (2003), 'A literature-based method for assessing the functional coherence of a gene group', *Bioinformatics*, Vol. 19(3), pp. 396–401.
51. Glenisson, P., Antal, P., Mathys, J. *et al.* (2003), 'Evaluation of the vector space representation in text-based gene clustering', in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 391–402.
52. Chiang, J. H. and Yu, H. C. (2003), 'MeKE: Discovering the functions of gene products from biomedical literature via sentence alignment', *Bioinformatics*, Vol. 19(11), pp. 1417–1422.
53. Raychaudhuri, S., Chang, J. T., Sutphin, P. D. and Altman, R. B. (2002), 'Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature', *Genome Res.*, Vol. 12(1), pp. 203–214.
54. Pan, H., Zuo, L., Choudhary, V. *et al.* (2004), 'Dragon TF Association Miner: A system for exploring transcription factor associations through text-mining', *Nucleic Acids Res.*, Vol. 32 (Web Server issue), pp. W230–234.
55. Krallinger, M. (2004), 'BioCreAtIvE – critical assessment of information extraction systems in

biology' (URL: http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html, accessed 1st September, 2004).

56. Eskin, E. and Agichtein, E. (2004), 'Combining text mining and sequence analysis to discover protein functional regions', in 'Proceedings of the 9th Pacific Symposium on Biocomputing', 6th–10th January, Hawaii, pp. 288–99.

57. Srinivasan, P. and Wedemeyer, M. (2003), 'Mining concept profiles with the vector model or where on earth are diseases being studied?', in 'Proceedings of the Text Mining Workshop. Third SIAM International Conference on Data Mining', San Francisco.

58. Kostoff, R. N. (2003), 'Bilateral asymmetry prediction', *Med. Hypotheses*, Vol. 61(2), pp. 265–266.

59. Xu, H., Anderson, K., Grann, V. and Friedman, C. (2004), 'Facilitating cancer research using natural language processing of pathology reports', in 'Proceedings of the 11th World Congress on Medical Informatics', IMIA, San Francisco, CA, pp. 565–569.

60. Swanson, D.R. (1991), 'Complementary structures in disjoint science literatures', in 'Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM Press, Chicago, IL, pp. 280–289.

61. Weeber, M., Vos, R., Klein, H. *et al.* (2003), 'Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide', *J. Amer. Med. Inform. Assoc.*, Vol. 10(3), pp. 252–259.

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

63. Ramadan, N. M., Halvorson, H., Vande-Linde, A. *et al.* (1989), 'Low brain magnesium in migraine', *Headache*, Vol. 29(9), pp. 590–593.

64. Ferrari, M. D. (1992), 'Biochemistry of migraine', *Pathol. Biol. (Paris)*, Vol. 40(4), pp. 287–292.

65. Gordon, M. D. and Lindsay, R. K. (1996), 'Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil', *J. Amer. Soc. Inf. Sci.*, Vol. 47(2), pp. 116–128.

66. Srinivasan, P. (2004), 'Text mining: Generating hypothesis from MEDLINE', *J. Amer. Soc. Inf. Sci. Technol.*, Vol. 55, pp. 396–413.

67. Srinivasan, P. and Libbus, B. (2004), 'Mining MEDLINE for implicit links between dietary substances and diseases', *Bioinformatics*, Vol. 20, Suppl. 1, pp. I290–296.

68. Srinivasan, P., Libbus, B. and Sehgal, A. K. (2004), 'Mining MEDLINE: Postulating a beneficial role for curcumin longa in retinal diseases', in 'BioLINK 2004: Linking Biological Literature, Ontologies, and Databases, Boston, MA, Association for Computational Linguistics, pp. 33–40 (URL: http:/www.cs.brandeis.edu/~jamesp/biolink2004/).

69. Novichkova, S., Egorov, S. and Daraselia, N. (2003), 'MedScan, a natural language processing engine for MEDLINE abstracts', *Bioinformatics*, Vol. 19(13), pp. 1699–1706.

70. Glenisson, P., Coessens, B., Van Vooren, S. *et al.* (2004), 'TXTGate: Profiling gene groups with text-based information', *Genome Biol.*, Vol. 5(6), p. R43.

71. Becker, K. G., Hosack, D. A., Dennis, G. Jr *et al.* (2003), 'PubMatrix: A tool for multiplex literature mining', *BMC Bioinformatics*, Vol. 4(1), p. 61.

72. Corney, D. P., Buxton, B. F., Langdon, W. B. and Jones, D. T. (2004), 'BioRAT: Extracting biological information from full-length papers', *Bioinformatics* (in press).

73. Müller, H., 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).

74. Nenadic, G., Spasic, I. and Ananiadou, S. (2003), 'Terminology-driven mining of biomedical literature', *Bioinformatics*, Vol. 19(8), pp. 938–943.

75. Aronson, A. R. (2001), 'Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program', in 'Proc. AMIA Symp.', 3rd–7th November, Washington, DC, pp. 17–21.

76. Rindflesch, T. C., Hunter, L. and Aronson, A. R. (1999), 'Mining molecular binding terminology from biomedical text', in 'Proc. AMIA Symp.', 6th–10th November, Washington, DC, pp. 127–131.

77. Majoros, W. H., Subramanian, G. M. and Yandell, M. D. (2003), 'Identification of key concepts in biomedical literature using a modified Markov heuristic', *Bioinformatics*, Vol. 19(3), pp. 402–407.

78. Regev, Y., Finkelstein-Landau, M. and Feldman, R. (2003), 'Rule-based extraction of experimental evidence in the biomedical domain: The KDD Cup 2002 (task 1)', *ACM SIGKDD Explorations Newsletter*, Vol. 4(2), pp. 90–92.

79. Hersh, W. R., Crabtree, M. K., Hickam, D. H. *et al.* (2002), 'Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions', *J. Amer. Med. Inform. Assoc.*, Vol. 9(3), pp. 283–293.