What the papers say: Text mining for genomics and systems biology

Nathan Harmston, Wendy Filsell** and Michael P.H. Stumpf *#*

*Division of Molecular Biosciences, Centre for Bioinformatics, Imperial College London, 303, Wolfson Building, South Kensington Campus, London, SW7 2AZ, UK

**Unilever R&D, Colworth Science Park, Sharnbrook, Bedford MK44 1 LQ, UK

#*Correspondence to:* Tel: +44 (0)20 7594 5114; Fax: +44 (0)20 7594 5789; E-mail: m.stumpf@imperial.ac.uk

Date received (in revised form): 6th August 2010

Abstract

Keeping up with the rapidly growing literature has become virtually impossible for most scientists. This can have dire consequences. First, we may waste research time and resources on reinventing the wheel simply because we can no longer maintain a reliable grasp on the published literature. Second, and perhaps more detrimental, judicious (or serendipitous) combination of knowledge from different scientific disciplines, which would require following disparate and distinct research literatures, is rapidly becoming impossible for even the most ardent readers of research publications. Text mining — the automated extraction of information from (electronically) published sources — could potentially fulfil an important role — but only if we know how to harness its strengths and overcome its weaknesses. As we do not expect that the rate at which scientific results are published will decrease, text mining tools are now becoming essential in order to cope with, and derive maximum benefit from, this information explosion. In genomics, this is particularly pressing as more and more rare disease-causing variants are found and need to be understood. Not being conversant with this technology may put scientists and biomedical regulators at a severe disadvantage. In this review, we introduce the basic concepts underlying modern text mining and its applications in genomics and systems biology. We hope that this review will serve three purposes: (i) to provide a timely and useful overview of the current status of this field, including a survey of present challenges; (ii) to enable researchers to decide how and when to apply text mining tools in their own research; and (iii) to highlight how the research communities in genomics and systems biology can help to make text mining from biomedical abstracts and texts more straightforward.

Keywords: *data mining, systems medicine, literature processing, hypothesis generation*

Introduction

The scientific literature provides an important source of knowledge generated by the research community; it does not become defunct five years after publication and it is not just something to promote the authors' careers. While large amounts of data relating to biological systems are stored in public repositories, an even larger amount can be found in a semi-structured form in the literature (see Figure 1). This knowledge is potentially very useful in a variety of genomics and systems biology contexts.[1] For example, manually curated and literature-derived protein-protein interaction datasets are typically used as gold standards by the systems biology community and it is standard practice to extract parameters for mechanistic models from the literature.

Manual curation lacks the scalability to deal with the ever-increasing numbers of papers being published[2,3] and suffers from inter-annotator disagreement: different curators may interpret a piece of text in different ways. This means that a single paper needs to be annotated at least twice if the

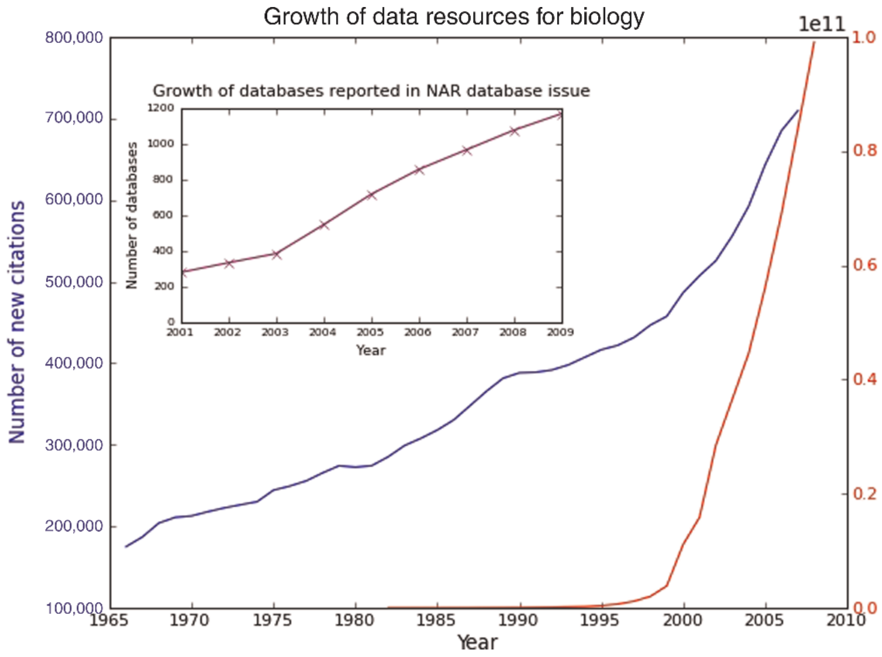

Figure 1. Biology is becoming a data-driven science, with an exponential growth in the number of papers being published, increasing numbers of databases indexed in the Nucleic Acids Research (NAR) database collection and an exponential growth in the number of base pairs stored in Genbank.

reliability of the proposed annotations is in any way to be calculated. The increase in the numbers of papers being published also means that it is becoming harder for researchers to stay up to date with the relevant literature in their field. This has an impact on their ability to generate meaningful and testable hypotheses, with some even suggesting that this is becoming a bottleneck in the scientific discovery process.[4]

These issues have motivated a sustained interest in the application of text mining (TM) techniques by both the industrial[5] and academic[6] communities to address some of these problems. TM refers to the process of extracting information encoded in text by authors through the use of techniques from a variety of fields such as information retrieval (IR), machine learning (ML), natural language processing (NLP), statistics and computational linguistics (CL).[7] The use of these techniques leads to a decrease in the time and effort required to extract information from a paper, speeding up curation[8] and also providing novel opportunities for hypothesis generation using the literature. We feel that, in the context of human genomics, this is particularly promising: an increasing number of studies report rare disease-causing variants and, in order to annotate such variants, assess their functional relevance or link them to existing clinical information, TM approaches will increase in importance as an enabling technology for biomedical research.

Text mining can be thought of as a method by which a systematic review can be performed. As with all methods for reviewing the existing literature, however, there are several biases. Due to copyright issues, only a relatively small number of papers are available for full-text mining and so most work is restricted to abstracts and titles, which are freely available from MEDLINE (only 30 per cent of curated protein–protein interactions (PPIs) can be found in the abstracts rather than the full text[9]). This does mean that extracted information is subject to a selection bias, although of a different form to that seen in manual curation (where only a subset of full text papers are curated). Neither manual curation nor TM techniques can deal with the inherent publication bias in the literature. Publication bias[10] refers to the fact that only positive findings (rather than negative or no findings) tend to get reported in the literature, and certain topics and/or genes tend to be reported more when they are in vogue. There is also evidence that PPI networks derived from the literature[11,12] are subject to ascertainment bias. This occurs when sampling is non-random and conclusions about the population are made based solely on this distorted group of frequently studied proteins. Conclusions about networks that are generated in the presence of ascertainment bias can dramatically change once the necessary corrections have been made.[13] Despite these biases, the literature is still extremely important to researchers as a method for communicating results and ideas and for testing and generating hypotheses.

Below, we give an overview of the current status of TM methodology. Some technical detail is required in order to appreciate fully the potential of this methodology, as well as its (current and future) limitations. At its worst, TM will be an exercise in high-throughput 'stamp collecting'; at best, it opens up the possibility of distilling vast amounts of published information into concrete hypotheses and functional insights into genomics and systems biology.

TM

In order to extract knowledge from text, named entities (NEs) must first be recognised; these NEs are then normalised to identifiers and any relationships between them are identified (see Figure 2). Biological NEs correspond to classes such as genes, proteins, cell lines, species, compounds, phenotypes, diseases, etc. Named entity recognition (NER) refers to the problem of labelling both the location (start, end) and the semantic class/type of a NE in text, and normalisation refers to the process of mapping a NE

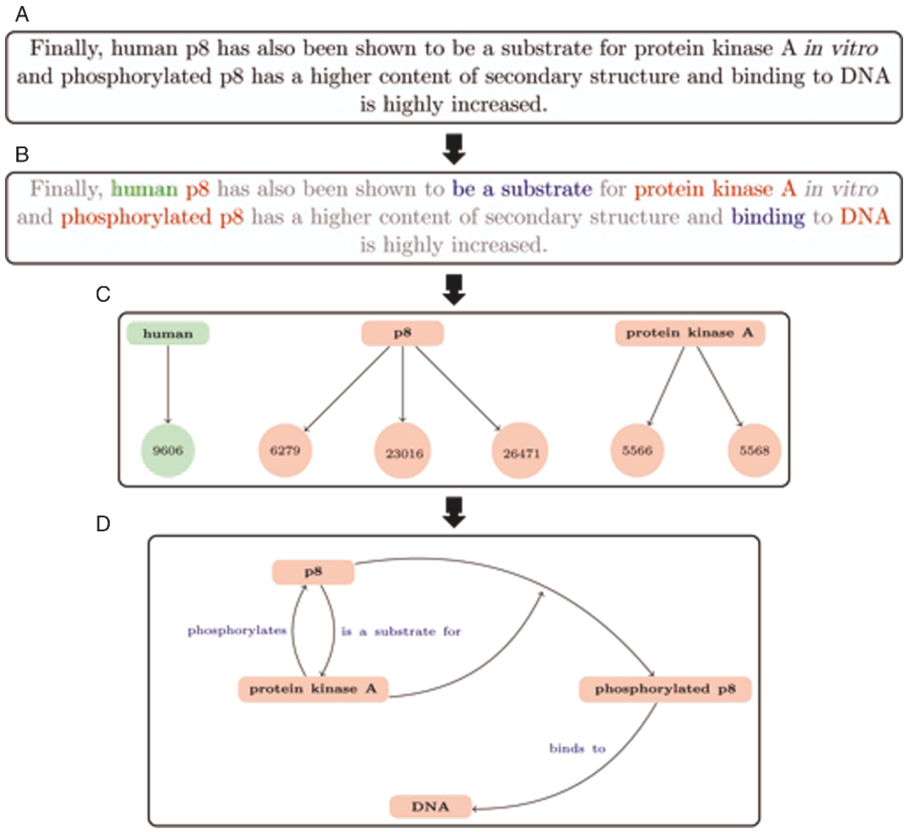
A
Finally, human p8 has also been shown to be a substrate for protein kinase A *in vitro* and phosphorylated p8 has a higher content of secondary structure and binding to DNA is highly increased.
B
Finally, human p8 has also been shown to be a substrate for protein kinase A *in vitro* and phosphorylated p8 has a higher content of secondary structure and binding to DNA is highly increased.
C
human
p8
protein kinase A
9606
6279
23016
26471
5566
5568
D
p8
phosphorylates
is a substrate for
protein kinase A
phosphorylated p8
binds to
DNA

Figure 2. An example of a text-mining pipeline. Given a sentence from a paper (A), named entities (NEs) are extracted (green for species entities, red for protein/gene entities, blue for relationship cues) (B); these entities are then normalised to a corresponding identifier scheme (C); and relationships between entities extracted (D). The final result in this case is a network which explicitly encodes the semantic relationships between NEs found in the sentence. Text taken from PMID:14613582.

to a unique identifier (or set of identifiers). Following NER and normalisation it is useful to determine if a real relationship exists between two or more NEs, as well as the type of relationship. Simply identifying that NEs occur together in a contiguous block of text does hint at the existence of some form of relationship; however, this relationship may be completely speculative, or the text may state that a relationship between the NEs does not exist.[14] In biological research papers, two entities can co-occur for many reasons, including functional, physical, syntenic and evolutionary relationships. The performance of TM systems is often evaluated using precision and recall metrics against manually curated gold standard corpora. Precision can be interpreted as the probability that a randomly selected result is a true positive and is calculated as the number of true positives obtained over the sum of true positives and false positives. Recall can be intuitively interpreted as the probability that a randomly selected positive result is correctly identified; it can be calculated as the number of true positives divided by the number of items that should be found (the sum of true positives and false negatives).

NER

Biology is a dynamic and ever-expanding research area. This means that there are millions of entity names in use, with new ones constantly being created (eg through genome annotation and drug development). Neologisms are prevalent in the literature; it has been jokingly commented: '*Scientists would rather share each other's underwear than use each other's nomenclature*' (Keith Yamamoto). Biological NER thus tends to be more difficult than NER tasks in other domains (eg newswires) due to the variability of biological nomenclatures.[15,16] A single gene can have many synonyms (eg *P53*, *TP53* and *TRP53* all refer to the same gene). Gene names are subject to morphological (eg transcription factor, transcriptional factor), orthographic (eg nuclear factor [NF] kappa B, NF κB), combinatorial (eg homologue of actin, actin homologue) and inflectional variation (eg antibody, antibodies). The HUGO Gene Nomenclature committee (HGNC) was created with the aim of assigning a unique gene symbol to every gene; however, currently, not all genes have been assigned a name and there are still problems with gene names mentioned in the past literature. Gene names can overlap with other names relating to different entity types in the biological domain, as well as with words that are found in everyday language. It is often difficult to disambiguate similar entity classes, as they can have similar contexts and morphologies. For example, a simple heuristic for determining whether a term refers to a gene or protein is that proteins begin with an upper case letter (*PspA*) and genes begin with a lower case (*pspA*). This pattern is, however, not maintained consistently in scientific writing, and humans show substantial disagreement on this task,[17] with an average pair-wise agreement among three annotators of 77.58 per cent.

The *Drosphilia melanogaster* literature is probably the best example of the problems that exist regarding nomenclatures. Some *Drosphilia* genes are named after their associated phenotype, such as *eyeless* or *fruity*, which leads to difficulties in disambiguating whether it is the phenotype or the gene that is being described. Gene names such as *Not* and *That* also exist, which are homonymous (see Table 1). Some gene names are multi-word names

Table 1. Table of linguistic terms. Definitions obtained from the Oxford Dictionary and WordNet

Term	Meaning
Anaphor	A word or phrase that refers back to an earlier word or phrase
Polysemy	The coexistence of many possible meanings for a word or phrase
Homonymy	Each of two or more words having the same spelling and pronunciation but different meanings and origins
Semantics	Relating the meaning in language or logic
Syntax	The arrangement of words and phrases to create well-formed sentences in a language
Part of speech	One of the traditional categories of words intended to reflect their functions in a grammatical context

Table 2. Some freely available software for NLP tasks in the biological domain. Task refers to the part of a text-mining pipeline that the software can be used for. Abbreviations: NER, named entity recognition; POS, part of speech tagger; PPI, protein–protein interaction extraction; SEN, sentencisation

Name	URL	Task
AbGene[28]	ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/AbGene	NER
ABNER[29]	http://pages.cs.wisc.edu/~bsettles/abner/	NER
AkanePPI[30]	http://www-tsujii.is.s.u-tokyo.ac.jp/~satre/akane/	PPI
BANNER[31]	http://banner.sourceforge.net/	NER
BioTagger-GM	http://www.seas.upenn.edu/~strctlrn/BioTagger/BioTagger.html	NER
GENIA[32]	http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/	NER
Graph Kernel[33]	http://mars.cs.utu.fi/PPICorpora/GraphKernel.html	PPI
LINNAEUS[34]	http://linnaeus.sourceforge.net/	NER
JNET[35]	http://julielab.de/	NER
JSBD	http://julielab.de/	SEN
LingPipe	http://alias-i.com/lingpipe/	NER
MedPOS[36]	http://ii-public.nlm.nih.gov/MMTx/MedPost_SKR.shtml	POS
NLProt[37]	http://cubic.bioc.columbia.edu/services/nlprot/	NER
OpenDMAP[38]	http://opendmap.sourceforge.net/	PPI
OSCAR3[39]	http://sourceforge.net/projects/oscar3-chem/	NER
POSBIOTM-NER[40]	http://isoft.postech.ac.kr/Research/BioNER/POSBIOTM/NER/main.html	NER
Sptoolkit	http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector	SEN
Whatizit[41]	http://www.ebi.ac.uk/webservices/whatizit/	NER/PPI

such as *Mind the gap* and *IL-2 receptor*. In the last case, problems detecting the correct boundary may lead to the entity being tagged as *IL-2*, which completely alters the meaning of the entity.[18]

A variety of methods have been proposed for biological NER (see Table 2), with only a small portion freely available for download or publicly accessible via web servers/services. These tools fall into four main categories: dictionary-based, rule-based/pattern-based, machine-learning and hybrid systems (and combinations of these approaches). Most research in this area has concentrated on recognising gene and protein mentions; however, there has also been some work on identifying cell lines, chemicals and species. Competitions such as NLPBA[19] and BioCreative[20] are held in order to evaluate NER methods for gene mention recognition.

Dictionary-based methods[21] work by matching text against a fixed dictionary of entity names. The performance of these methods is highly dependent on both the coverage of the dictionary and the performance of matching techniques used. Use of a simple text-matching algorithm will lead to a large number of false positives being found because of the overlap between dictionary words and common English, as well as some false negatives due to misspellings not present in the dictionary. Gene names which lead to false positives are typically filtered out of dictionaries. Most systems that are based on this method either use an approximate method of string matching[22] or expand the dictionary by generating spelling variants.[23,24] These methods tend to lead to an increase in recall accompanied by a decrease in precision. In some cases, dictionary-based NER methods can perform normalisation at the same time.[25]

Rule-based methods[26] use orthographic and morpho-syntactic features of NEs (capital letters, numbers, symbols and affixes) and their surrounding words to generate patterns and rules. Biochemical suffixes such as -*ase* and -*in* are very useful in indicating possible protein names and so a simple rule would be to tag words with these features as proteins. These systems incorporate expert knowledge easily and the rules generated are human readable and easily extendable. Rule-based techniques are able to reach high levels of precision but at the expense of recall, as they are not robust against unseen names. This is mainly because there are so many potential surface grammatical variations (active, passive voice) and it is not feasible to develop robust patterns for all of these.

Machine learning (ML) methods tend to achieve the highest performance for NER. All of the top ten performing methods in the BioCreative II gene mention task (BCII GM) used a machine-learning component. ML methods use training data in the form of a manually annotated gold standard corpus and learn features that are useful in identifying NEs in text. The performance of the methods used in NER can be very sensitive to feature selection, although this is not always the case.[27] NER can be viewed as either a classification or a sequence-labelling problem. Classification approaches normally consider NER as assigning a class to a bag of features. These features include surface clues and morpho-syntactic features of NEs and their adjacent words. These methods do not tend to take the order of features into account and support only binary classifications. Sequence labelling approaches deduce the most probable sequence of tags for a given sequence of words. Each token is assigned a tag by calculating the most likely label for the current token, given both the features of that token and the previous history of tag assignments. The performance of any ML tagger will be biased by the size, inter-annotator agreement and topic structure of the corpus (see Table 3).

Determining the correct class of an NE is complicated by the ubiquitous use of abbreviations and acronyms in biomedical research. Liu *et al.*[42] found that 81.2 per cent of acronyms in MEDLINE are ambiguous (eg the acronym NF can refer to 61 different full forms[43]). ML methods have been proposed for abbreviation disambiguation,[44] with some work focusing on abbreviations found in the biological literature.[43,45]

Table 3. Freely available corpora for training and evaluating text-mining tools in the biological domain. Task refers to the tool training/evaluation use of the corpus. Abbreviations: GM, gene mention (NER); GN, gene normalisation; REL, relationship extraction; SD, species; SM, species mention (NER); SN, species normalisation

Corpus	Location	Task
AIMed	ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/	REL
BioCreative I GM	http://sourceforge.net/projects/biocreative/files/	GM
BioCreative Ib	http://sourceforge.net/projects/biocreative/files/	GN
BioCreative II GM	http://sourceforge.net/projects/biocreative/files/	GM
BioCreative II GN	http://sourceforge.net/projects/biocreative/files/	GN
BioCreative II FT	http://sourceforge.net/projects/biocreative/files/	REL
BioInfer [48]	http://www.it.utu.fi/BioInfer/	REL
DECA [49]	http://www.nactem.ac.uk/deca_details/start.cgi	SD
LINNAEAUS [34]	http://linnaeus.sourceforge.net/	SM/SN

It is not just gene names that are difficult to identify; the identification of species mentions is also troublesome. Species names can be homonymous with common English words (eg 'honesty' for *Lunaria annua* and 'bears' for *Ursidae*) but also with important entities in the biological domain (eg *cancer* and *hippocampus*). The performance of a dictionary-based tagging system is again limited by the lack of coverage, widespread use of acronyms and frequent misspelling of species names. Standard dictionaries of species names such as the National Center for Biotechnology Information (NCBI) Taxonomy are incomplete, given the amazing diversity of life. They do, however, contain names for most well-studied organisms. Rule-based methods[46] have been developed which are capable of identifying species terms using rules designed for matching Linnaean binomial nomenclature. The recently published LINNAEAUS[34] system uses a dictionary and a set of regular expressions to identify species mentions in text. This system allows both the identification and normalisation of species names, features an acronym disambiguation component and achieves high performance on its own corpus.

Cell lines are widely used in biological and biomedical research as a platform for functional studies and to validate biomarkers. It is useful to identify cell line mentions as they can aid in identifying experimental techniques/conditions and to determine the species to which other entity types belong during normalisation. A recent analysis of the cell line nomenclature[47] revealed that it, too, is blighted by ambiguity and variability. Several NER taggers have been trained to identify cell line mentions in text, although there is not yet one specifically designed for tagging cell line mentions. Recently, integrating information from different sources has led to the creation of a cell line knowledge base (CLKB). This work represents the start of efforts to create a lexicon of cell line names, although it is incomplete, so dictionary-based techniques may still miss cell line mentions. As with other subsets of biological nomenclature, there is vertical polysemy (see Table 1) with other NE classes (see Figure 3).

Entity normalisation

Normalisation of NEs allows the results of text mining to be used in tasks like manual curation,[50] knowledge summarisation[51] and model construction and validation.[52,53] The standard method of normalisation is to compare an NE against a dictionary of synonyms and identifiers, and assign the matching identifier. In some domains, this approach can achieve an extremely good performance; however, the variability and ambiguity of biological nomenclature means that this method is essentially ineffective for biological entities. The genomic nomenclature is

Figure 3. (A) **HU**man **N**atural **K**iller; (B) Large piece of something without definite shape; (C) A well-built, sexually attractive man; (D) **H**ormonally **U**pregulated **N**eu-associated **K**inase. Demonstration of the possible problems due to the biological nomenclature, given the sentence *HUNK is associated with expression of Frizzled-2*: HUNK could refer to a cell type, a protein and two common English words. While, in biological text, it is highly probable that (B) and (C) will not be relevant, it is not so easy to disambiguate (A) and (D). This is an example of the problems posed by polysemy (a word or phrase having multiple meanings), homonymity with common English words and the use of abbreviations in the literature.[18]

also highly ambiguous, in that one gene name can map to multiple canonical identifiers. This means that exact text matching using a dictionary is flawed, as the term may be a variation not found in the list of synonyms. Rule-based approaches[54] have been used which try to normalise terms by applying a set of transformations to a tagged entity in order to try to make it match a term in a lexicon. String similarity metrics[55] have been used with some success[56] to match terms which are not present in the original lexicon.

Due to the ambiguity in biological nomenclatures (Figure 4), it is important to disambiguate between multiple identifiers. Several approaches have been proposed in order to deal with this problem: rule-based, ML or hybrid. Rule-based approaches[57] use various heuristics to try to assign scores to identifiers. The creation of bags of words associated with specific identifiers (known as semantic profiles) has been useful for disambiguation. These profiles are created by extracting information from various genomic knowledge sources such as UniProt, GO and Entrez. These can then be used to train a classifier to distinguish the correct identifier from incorrect ones.[58] Knowledge of paper co-authorship has been found to be useful in identifier disambiguation,[59] based on the idea that an author uses gene names consistently across all of their publications or may work on a specific set of genes consistently.

It is not just the proteomic and genomic nomenclatures that pose problems for normalisation. While the precise Linnaean binomial name for an organism is unambiguous, it may not be the case for its abbreviated form. *Caenorhabditis elegans* is commonly abbreviated to *C. elegans*; however, 49 other species have a name that can be abbreviated to this short form. Due to the widespread use of *Caenorhabditis elegans* as a model organism, the majority of mentions of *C. elegans* would probably normalise to NCBI Taxonomy identifier 6239 but this heuristic will have exceptions. Another problem with species normalisation is dealing with the abundance of different strains, particularly among microorganisms. It is important to

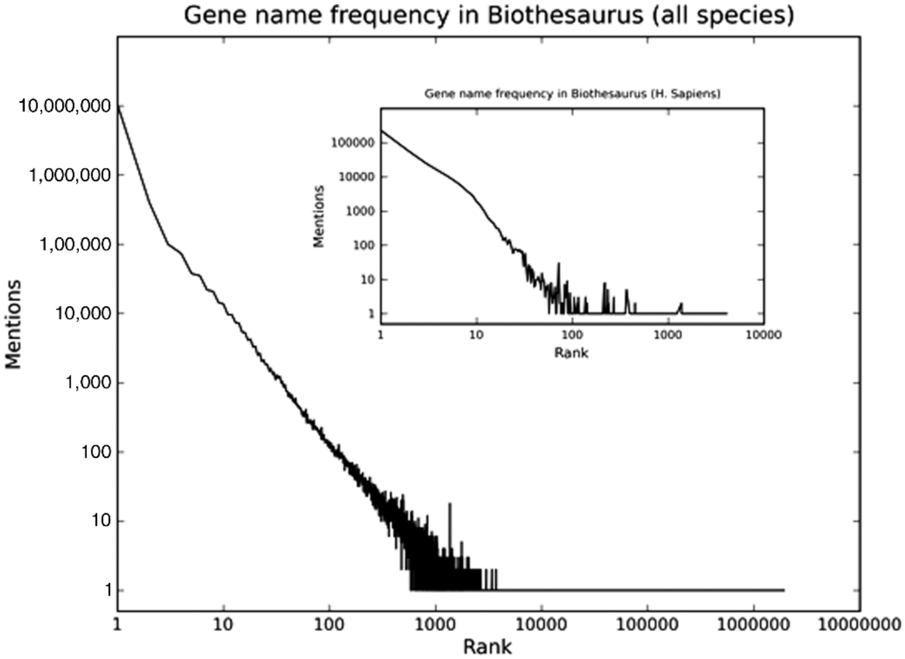

Figure 4. The genomic nomenclature is highly ambiguous. The plot shows the rank of a gene name against the total number of times that the gene name is found in Biothesaurus. The inset shows this only looking at human genes. The plot is in log–log coordinates. Both graphs show Zipf-like (discrete power-law) distributions. Biothesaurus is a collection of gene names mapped to Entrez Gene/ Uniprot identifiers across approximately 7,000 species.

disambiguate the strains if possible, as genes' functional properties can vary between strains.

Good results for normalising human gene names have been reported. The BCII GN task[60] evaluated performance against a manually annotated gold standard corpus. Overall results were promising, with a combined recall of 97.2 per cent (entries from over 20 teams). This evaluation assumed that the species was human, however. Normalisation for other species continues to be a challenge and has not been helped by the decision made at the 22nd International Society for Animal Genetics (in August 1990) that animal gene names should 'follow the rules for human gene nomenclature, including the use of identical symbols for homologous genes and the reservation of human symbols for as yet unidentified animal genes'.[15] This interspecies ambiguity of the genomic nomenclature means that identifying the correct species for a given mention is an important subtask of gene normalisation, although it has only recently begun to be considered.[61]

Relation extraction

Identifying the existence and type of relationships between entities is difficult because of the numerous ways that a relationship can be proposed. A binding relationship between two proteins could, for example, be described in at least three ways:

(1) APPL binds Akt2
(2) Binding of Akt2 by APPL
(3) Binding between Akt2 and APPL

Relationships between two entities can be described over multiple sentences, which can lead to complications, as anaphors need to be identified and resolved (eg *APPL* is later referred to as *this protein* in a piece of text). This limits the recall of relation extraction approaches that work at the sentence level only. The relationship type that has attracted the most effort is extracting PPIs.

A number of different approaches have been proposed in order to perform this task based on linguistic, rule-based and ML methods. Rule-based methods use a set of syntactic patterns, which specify how an interaction is described. The patterns can be manually or automatically generated. RelEX[62] applies a simple set of rules on a representation of the dependencies between words in a sentence called a dependency graph. The RLIMS-P[63] is a rule-based approach specifically designed to extract information about protein phosphorylation sites, and performs well compared with manually curated literature sets. Some ML methods treat a sentence as a sequence of words or tokens and completely ignore its syntactic structure. These approaches do not achieve good performance compared with methods which take sentence structure into account. It is clearly important to consider both contextual and linguistic features,[64,65] such as interaction keywords and verbs,[66] to extract relationships with good precision.

To complicate matters further, authors frequently speculate about potential relationships (eg APPL may interact with Akt2). These statements do not correspond to the definition of a relationship, but that the relationship is proposed to exist. It is important to identify these speculative statements[67] and prevent them from biasing any downstream analyses. For the same reason, it is equally important to detect the negation of relationships[68] (eg APPL does not interact with Akt2).

Hypothesis generation

The scientific literature not only contains explicit knowledge, such as 'APPL interacts with Akt2', but also implicit knowledge,[69] such as hidden refutations or qualifications, inferences from transitive relations, hidden or unrecognised analogies and the accumulation of weak tests (which could be used in meta-analyses). Swanson's serendipitous discovery of the connection between Raynaud's disease and fish oil[70] is an example of performing an inference on a transitive relation to generate a novel and testable hypothesis. By reading two disjoint sets of literature (no articles are in common, and the articles in one set do not cite or mention articles in the other set), he observed that blood factors were a common theme in both the Raynaud's disease and the fish-oil literature. This led him to propose that fish oil could be used in the treatment of Raynaud's disease, and the relationship was clinically validated in 1989.[71] The discovery led Swanson to propose that 'new hypotheses can emerge and scientific discovery can be anticipated or stimulated through the investigation of complementary but disjoint literatures'. This method of literature-based discovery is commonly referred to as Swanson's ABC model or Swanson Linking, with the hypotheses and new knowledge being described as undiscovered public knowledge. Although the model has mainly been used within the biomedical and biological fields it has also been applied to the humanities literature and the WWW (see Table 4).

Mendeleev's discovery of the law of periodicity and the development of the periodic table can be considered an early example of literature-based discovery (LBD), as it was: 'a direct outcome of the stock of generalisations and established facts which has accumulated by the end of the decade 1860–1870.' The information required to build the table of elements had already been published, but it had never been analysed as a whole.[72] More recently, Hettne *et al.*[73] combined TM with network analysis in order to generate new mechanistic hypotheses relating to the complex regional pain syndrome (CRPS). NF-κB was identified as potentially being involved by first extracting genes relating to CRPS from the literature and then investigating potential links between these genes which were not mentioned in the CRPS literature. This hypothesis has led to several new ideas regarding the aetiology of the disease and the proposal of a novel drug target. By exploiting the context of protein mentions, van Haagen *et al.*[74] were able to predict a novel interaction between CAPN3 and PARVB. Integrating information extracted from the literature with microarray experiments has led to the proposition of a relationship between SIP and the invasiveness of glioblastoma cell lines.[75] All of this work shows the potential for TM to generate testable hypotheses for use in biology.

Hypothesis generation is challenging even to humans, however. Automating this process, or formulating it in such a way that a computer can

Table 4. Summary of hypotheses generated using Swanson's ABC model and its extensions

Paper	Hypothesis
Cory *et al.*[76]	Proposed links between Frost (a 20th century poet) and Carneades (an ancient philosopher)
Gordon *et al.*[77]	Finding new applications for genetic algorithms using the WWW
Hettne *et al.*[73]	Proposed the role of NF-κb in the aetiology of complex regional pain syndrome
Hristovski *et al.*[78]	Proposed novel candidate genes that may be involved in bilateral perisylvian polymicrogyria
Kostoff *et al.*[79]	Proposed novel non-drug treatments (such as calorific restriction) for the treatment of multiple sclerosis
Kostoff *et al.*[80]	Proposed 'lifestyle/dietary practices that could be interpreted as anti-cataract'
Srinivasan *et al.*[81,82]	Novel uses for curcuma longa/turmeric in the treatment of retinal diseases, Crohn's disease and spinal cord-related disorders
Swanson *et al.*[83]	Classifying viruses as potential biological weapons
van Haagen *et al.*[74]	Predicting and identifying novel interaction partners for proteins in *Escherichia coli*
Weeber *et al.*[84]	Novel uses of thalidomide in the treatment of myasthenia gravis, chronic hepatitis C, *Heliobacter pylori*-induced gastritis and acute pancreatitis
Wren *et al.*[85]	Chlorpromazine may reduce cardiac hypertrophy (ABC model in conjunction with experimental evidence)
Wren *et al.*[86]	Pathogenesis of non-insulin-dependent diabetes is most likely epigenetic
Zhou *et al.*[87]	Combined MEDLINE with traditional Chinese medicine to propose new functional knowledge about genes

quickly generate testable scientific propositions, is a non-trivial and daunting task. Only if the universe of potential hypotheses is sufficiently simple for search or enumeration approaches to cover all potential cases is this currently feasible. We feel that the most promising strategies in the short term include the search for suitable heuristics or iterative procedures involving infrequent human input.

Conclusion

TM tools offer a way to retrieve the pertinent information contained within the mass of scientific literature, make it easier to explore[88] and allow the generation of novel insights into existing data, all in an automated fashion. While TM is currently noisy and imperfect, it should be remembered that, due to inter-annotator disagreement, manual curation is too. TM is not just restricted to extracting functional information; it has also been used to identify best practices within the phylogenetics domain,[89] to generate priors for network reconstruction using Bayesian networks[90] and to aid in protein structure comparison and assignment of function.[91] Recently, TM has shown the greatest potential when used in data fusion style approaches. By using information extracted from the literature, Raychaudhuri *et al.*[92] were able to develop a method better to distinguish between genomic regions associated with disease and false-positive regions. Ten out of 13 single nucleotide polymorphisms (SNPs) identified by their method as been associated with Crohn's disease were later validated by follow-up genotyping. STRING[93] integrates many different types of evidence about PPIs, including literature co-occurrence, phylogenetic data and results from high-throughput experiments, and has been used to predict novel PPIs in other organisms by transferring annotations to orthologous protein pairs. While there is a significant body of work on applying TM to the biological domain, however, there still remain many challenges in areas like relation extraction, species disambiguation and hypothesis generation.

Systems biology and genomics deal with large data models of unprecedented complexity; TM allows us to draw on the published literature in a disciplined manner to inform the development of quantitative models. We expect TM to become an important addition to the systems biologist's toolkit, complementing existing techniques like comparative and primary data analysis. We hope to have demonstrated the use and limitations of TM in its current guise. Being aware of the limitations, however, should enable the community to develop and adopt protocols that allow for easier, more reliable analysis of published research outputs from these tools. This is important not only for researchers, but also for publishers, funding bodies and regulators. These three players have, of course, different but, crucially, not competing interests as far as accessibility of information is concerned. Regulators, in particular, irrespective of whether or not they are engaged in accrediting new drugs or nutritional supplements or the granting of patents, stand to benefit profoundly from information that is provided in an electronically accessible and unambiguous fashion.

References

1. Ananiadou, S., Kell, D. and Tsujii, J. (2006), 'Text mining and its potential applications in systems biology', *Trends Biotechnol*. Vol. 24, pp. 571–579.
2. Baumgartner, W.A., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G. *et al*. (2007), 'Manual curation is not sufficient for annotation of genomic databases', *Bioinformatics* Vol. 23, pp. i41–i48.
3. Winnenburg, R., Wächter, T., Plake, C., Doms, A. *et al*. (2008), 'Facts from text: Can text mining help to scale-up high-quality manual curation of gene products with ontologies?', *Brief. Bioinform*. Vol. 9, pp. 466–478.
4. Ng, S. and Wong, M. (1999), 'Toward routine automatic pathway discovery from on-line scientific text abstracts', *Genome Inform*. Vol. 10, pp. 104–112.
5. Agarwal, P. and Searls, D.B. (2008), 'Literature mining in support of drug discovery', *Brief. Bioinform*. Vol. 9, pp. 479–492.
6. Rzhetsky, A., Seringhaus, M. and Gerstein, M. (2008), 'Seeking a new biology through text mining', *Cell* Vol. 134, pp. 9–13.
7. Hearst, M. (1999), 'Untangling text data mining', Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics', pp. 3–10.
8. Deshpande, N., Fink, J., Bourne, P. and Cohen, K. (2008), 'Intrinsic evaluation of text mining tools may not predict performance on realistic tasks', Pacific Symposium on Biocomputing, pp. 640–651.
9. Blaschke, C. (2001), 'Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study', *Comp. Funct. Genomics* Vol. 2, pp. 196–206.
10. Knight, J. (2003), 'Negative results: Null and void', *Nature* Vol. 422, pp. 554–555.
11. Pfeiffer, T. and Hoffmann, R. (2003), 'Temporal patterns of genes in scientific publications', *Proc. Natl Acad. Sci. USA* Vol. 104, pp. 12052–12056.

12. Lehne, B. and Schlitt, T. (2009), 'Protein-protein interaction databases: Keeping up with growing interactomes', *Hum. Genomics* Vol. 3, pp. 291–297.
13. Dickerson, J., Pinney, J. and Robertson, D. (2010), 'The biological context of HIV-1 host interactions reveals subtle insights into a system hijack', *BMC Syst. Biol.* Vol. 4, p. 80.
14. Jenssen, T., Lægreid, A., Komorowski, J. and Hovig, E. (2001), 'A literature network of human genes for high-throughput analysis of gene expression', *Nat. Genet.* Vol. 28, pp. 21–28.
15. Chen, L., Liu, H. and Friedman, C. (2005), 'Gene name ambiguity of eukaryotic nomenclatures', *Bioinformatics* Vol. 21, pp. 248–256.
16. Mons, B. (2005), 'Which gene did you mean?', *BMC Bioinform.* Vol. 6, p. 142.
17. Hatzivassiloglou, V., Duboue, P.A. and Rzhetsky, A. (2001), 'Disambiguating proteins, genes, and RNA in text: a machine learning approach'. *Bioinformatics* Vol. 17, pp. 97–106.
18. Barnes, J. (2002), 'Conceptual biology: A semantic issue and more', *Nature* Vol. 417, pp. 587–588.
19. Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y.N. *et al.* (2004), 'Introduction to the bio-entity recognition task at JNLPBA', Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 70–75.
20. Smith, L., Tanabe, L.K., Johnson, R., Kuo, C.J. *et al.* (2008), 'Overview of BioCreative II gene mention recognition', *Genome Biol.* Vol. 9, pp. S2.
21. Liu, H., Hu, Z.Z., Torii, M., Wu, C. *et al.* (2006), 'Quantitative assessment of dictionary-based protein named entity tagging', *J. Am. Med. Inform. Assoc.* Vol. 13, pp. 497–507.
22. Tsuruoka, Y., McNaught, J., Tsujii, J. and Ananiadou, S. (2007), 'Learning string similarity measures for gene/protein name dictionary look-up using logistic regression', *Bioinformatics* Vol. 23, pp. 2768–2774.
23. Schuemie, M., Mons, B., Weeber, M. and Kors, J. (2007), 'Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification', *J. Biomed. Inform.* Vol. 40, pp. 316–324.
24. Tsuruoka, Y. (2003), 'Probabilistic term variant generator for biomedical terms', Proceedings of the 26th Annual International ACM SIGR Conference on Research and Development in Information Retrieval, pp. 167–173.
25. Fundel, K., Güttler, D., Zimmer, R. and Apostolakis, J. (2005), 'A simple approach for protein name identification: Prospects and limits', *BMC Bioinform.* Vol. 6 (Suppl. 1), p. S15.
26. 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, pp. 135–143.
27. Hakenberg, J., Bickel, S., Plake, C., Brefeld, U. *et al.* (2005), 'Systematic feature evaluation for gene name recognition', *BMC Bioinform.* Vol. 6, pp. S9.
28. Tanabe, L. and Wilbur, W. (2002), 'Tagging gene and protein names in biomedical text', *Bioinformatics* Vol. 18, pp. 1124–1132.
29. Settles, B. (2005), 'ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text', *Bioinformatics* Vol. 21, pp. 3191–3192.
30. Sætre, R., Sagae, K. and Tsujii, J. (2007), 'Syntactic features for protein-protein interaction extraction', Proceedings of the 2nd International Symposium on Languages in Biology and Medicine, pp 6.1–6.14.
31. Leaman, R. and Gonzalez, G. (2008), 'BANNER: An executable survey of advances in biomedical named entity recognition', Pacific Symposium on Biocomputing, pp. 652–663.
32. Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T. *et al.* (2005), 'Developing a robust part-of-speech tagger for biomedical text', Proceedings of Panhellenic Conference on Informatics, Vol. 3746, pp. 382–392.
33. Airola, A., Pyysalo, S., Björne, J., Pahikkala, T. *et al.* (2008), 'All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning', *BMC Bioinform.* Vol. 9, p. S2.
34. Gerner, M., Nenadic, G. and Bergman, C.M. (2010), 'LINNAEUS: A species name identification system for biomedical literature', *BMC Bioinform.* Vol. 11, p. 85.
35. Hahn, U., Buyko, E. and Landefeld, R. (2008), 'An overview of JCoRe, the JULIE lab UIMA component repository', Proceedings of the LREC'08 Workshop Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, pp. 1–7.
36. Smith, L., Rindflesch, T. and Wilbur, W. (2004), 'MedPost: A part-of-speech tagger for bioMedical text', *Bioinformatics* Vol. 20, pp. 2320–2321.
37. Mika, S. and Rost, B. (2004), 'NLProt: Extracting protein names and sequences from papers', *Nucleic Acids Res.* Vol. 32, pp. W634–W637.
38. Hunter, L., Lu, Z., Firby, J., Baumgartner, W.A. *et al.* (2008), 'OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression', *BMC Bioinform.* Vol. 9, p. 78.
39. Corbett, P. and Murray-Rust, P. (2006), 'High-throughput identification of chemistry in life science texts', Proceedings of the 2nd International Symposium on Computational Life Science, pp 107–118.
40. Song, Y., Kim, E., Lee, G.G. and Yi, B.K. (2005), 'POSBIOTM-NER: A trainable biomedical named-entity recognition system', *Bioinformatics* Vol. 21, pp. 2794–2796.
41. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. *et al.* (2008), 'Text processing through Web services: calling Whatizit', *Bioinformatics* Vol. 24, pp. 296–298.
42. Liu, H., Aronson, A.R. and Friedman, C. (2002), 'A study of abbreviations in MEDLINE abstracts'. Proceedings/AMIA Annual Symposium AMIA Symposium, pp. 464–468.
43. Okazaki, N. and Ananiadou, S. (2006), 'Building an abbreviation dictionary using a term recognition approach', *Bioinformatics* Vol. 22, pp. 3089–3095.
44. Tsuruoka, Y. and Ananiadou, S. (2005), 'A machine learning approach to acronym generation'. Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, pp. 25–31.
45. Bracewell, D., Russell, S. and Wu, A. (2005), 'Identification, expansion, and disambiguation of acronyms in biomedical texts', *Lect. Notes Comput. Sci.* Vol. 3759, pp. 186–195.
46. Koning, D., Sarkar, I. and Moritz, T. (2005), 'TaxonGrab: Extracting taxonomic names from text', *Biodiversity Inform.* Vol. 2, pp. 79–82.
47. Sarntivijai, S., Ade, A.S., Athey, B.D. and States, D.J. (2008), 'A bioinformatics analysis of the cell line nomenclature', *Bioinformatics* Vol. 24, pp. 2760–2766.
48. Pyysalo, S., Ginter, F., Heimonen, J., Björne, J. *et al.* (2007), 'BioInfer: A corpus for information extraction in the biomedical domain', *BMC Bioinform.* Vol. 8, p. 50.
49. Wang, X., Tsujii, J. and Ananiadou, S. (2010), 'Disambiguating the species of biomedical named entities using natural language parsers', *Bioinformatics* Vol. 26, pp. 661–667.
50. Alex, B., Grover, C., Haddow, B., Kabadjov, M. *et al.* (2008), 'Assisted curation: Does text mining really help?', *Pac. Symp. Biocomput.* pp. 556–567.
51. Craven, M. and Kumlien, J. (1999), 'Constructing biological knowledge bases by extracting information from text sources', Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pp. 77–86.
52. Santos, C., Eggle, D. and States, D. (2005), 'Wnt pathway curation using automated natural language processing: Combining statistical methods with partial and full parse for knowledge extraction', *Bioinformatics* Vol. 21, pp. 1653–1658.
53. Waagmeester, A., Pezik, P., Coort, S., Tourniaire, F. *et al.* (2009), 'Pathway enrichment based on text mining and its validation on carotenoid and vitamin A metabolism', *OMICS* Vol. 13, pp. 367–379.
54. Lau, W.W., Johnson, C.A. and Becker, K.G. (2007), 'Rule-based human gene normalization in biomedical text with confidence estimation', *Comput. Syst. Bioinformatics Conf.* Vol. 6, pp. 371–379.
55. Wang, X. and Matthews, M. (2008), 'Comparing usability of matching techniques for normalising biomedical named entities', *Pac. Symp. Biocomput.* Vol. 13, pp. 628–639.

56. Grover, C., Haddow, B., Klein, E. and Matthews, M. (2007), 'Adapting a relation extraction pipeline for the BioCreAtIvE II task', Proceedings of the BioCreAtIvE II Workshop.
57. Wang, X. (2007), 'Rule-based protein term identification with help from automatic species tagging', Proceedings of CICLING, pp. 288–298.
58. Crim, J., McDonald, R. and Pereira, F. (2005), 'Automatically annotating documents with normalized gene lists', *BMC Bioinform*. Vol. 6, p. S13.
59. Farkas, R. (2008), 'The strength of co-authorship in gene name disambiguation', *BMC Bioinform*. Vol. 9, p. 69.
60. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M. *et al*. (2008), 'Overview of BioCreative II gene normalization', *Genome Biol*. Vol. 9, p. S3.
61. Kappeler, T., Kaljurand, K. and Rinaldi, F. (2009), 'TX task: Automatic detection of focus organisms in biomedical publications', Proceedings of the Workshop on BioNLP, pp. 80–88.
62. Fundel, K., Kuffner, R. and Zimmer, R. (2007), 'RelEx–Relation extraction using dependency parse trees', *Bioinformatics* Vol. 23, pp. 365–371.
63. Hu, Z.Z., Narayanaswamy, M., Ravikumar, K.E., Vijay-Shanker, K. *et al*. (2005), 'Literature mining and database annotation of protein phosphorylation using a rule-based system', *Bioinformatics* Vol. 21, pp. 2759–2765.
64. Niu, Y., Otasek, D. and Jurisica, I. (2009), 'Evaluation of linguistic features useful in extraction of interactions from PubMed: Application to annotating known, high-throughput and predicted interactions in I2D', *Bioinformatics* Vol. 26, pp. 111–119.
65. Fayruzov, T., Cock, M.D., Cornelis, C. and Hoste, V. (2009), 'Linguistic feature analysis for protein interaction extraction', *BMC Bioinform*. Vol. 10, p. 374.
66. Hatzivassiloglou, V. and Weng, W. (2002), 'Learning anchor verbs for biological interaction patterns from published text articles', *Int. J. Med. Inform*. Vol. 67, pp. 19–32.
67. Kilicoglu, H. and Bergler, S. (2008), 'Recognizing speculative language in biomedical research articles: A linguistically motivated perspective', *BMC Bioinform*. Vol. 9, p. S10.
68. Sanchez-Graillet, O. and Poesio, M. (2007), 'Negation of protein-protein interactions: Analysis and extraction', *Bioinformatics* Vol. 23, pp. i424–i432.
69. Davies, R. (1989), 'The creation of new knowledge by information retrieval and classification', *J. Doc*. Vol. 4, pp. 273–301.
70. Swanson, D. (1986), 'Fish oil, Raynaud's syndrome, and undiscovered public knowledge', *Perspect. Biol. Med*. Vol. 30, pp. 7–18.
71. DiGiacomo, R., Kremer, J. and Shah, D. (1989), 'Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study', *Am. J. Med*. Vol. 86, pp. 158–164.
72. Murray-Rust, P. (2007), 'Data Driven Science–A Scientist's View', NSF/JISC 2007 Digital Repositories Workshop, Available at http://www.sis.pitt.edu/~repwkshop/papers/murray.html.
73. Hettne, K., de Mos, M., de Bruijn, A. and Weeber, M. (2007), 'Applied information retrieval and multidisciplinary research: New mechanistic hypotheses in complex regional pain syndrome', *J. Biomed. Discov. Collab*. Vol. 2, p. 2.
74. van Haagen, H., 't Hoen, P., Bovo, A.B., de Morrée, A. *et al*. (2009), 'Novel protein-protein interactions inferred from literature context', *PLoS ONE* Vol. 4, p. e7894.
75. Natarajan, J., Berrar, D., Dubitzky, W., Hack, C. *et al*. (2006), 'Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line', *BMC Bioinform*. Vol. 7, p. 373.
76. Cory, K. (1997), 'Discovering hidden analogies in an online humanities database', *Comput. Hum*. Vol. 31, pp. 1–12.
77. Gordon, M., Lindsay, R. and Fan, W. (2002), 'Literature-based discovery on the World Wide Web', *ACM Trans. Inter. Tech*. Vol. 2, pp. 261–275.
78. Hristovski, D., Peterlin, B., Mitchell, J and Humphrey, S. (2005), 'Using literature-based discovery to identify disease candidate genes', *Int. J. Med. Inform*. Vol. 74, pp. 289–298.
79. Kostoff, R., Briggs, M. and Lyons, T. (2007), 'Literature-related discovery (LRD): Potential treatments for multiple sclerosis', *Technol. Forecast. Soc. Change* Vol. 75, pp. 239–255.
80. Kostoff, R. (2007), 'Literature-related discovery (LRD): Potential treatments for cataracts', *Technol. Forecast. Soc. Change* Vol. 75, pp. 215–225.
81. Srinivasan, P. and Libbus, B. (2004), 'Mining MEDLINE for implicit links between dietary substances and diseases'. *Bioinformatics* Vol. 20, pp. i290–i296.
82. Srinivasan, P., Libbus, B. and Sehgal, A. (2004), 'Mining medline: Postulating a beneficial role for curcumin longa in retinal diseases', HLT Biolink, pp. 33–40.
83. Swanson, D., Smalheiser, N. and Bookstein, A. (2001), 'Information discovery from complementary literatures: Categorizing viruses as potential weapons', *J. Am. Soc. Inf. Sci. Technol*. Vol. 52, pp. 797–812.
84. Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, LT *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. Am. Med. Inform. Assoc*. Vol. 10, pp. 252–259.
85. Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V. *et al*. (2004), 'Knowledge discovery by automated identification and ranking of implicit relationships', *Bioinformatics* Vol. 20, pp. 389–398.
86. Wren, J. (2005), 'Data-mining analysis suggests an epigenetic pathogenesis for type 2 diabetes', *J. Biomed. Biotechnol*. Vol. 2, pp. 104–112.
87. Zhou, X., Liu, B., Wu, Z. and Feng, Y. (2007), 'Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks', *Artif. Intell. Med*. Vol. 41, pp. 87–104.
88. Hoffmann, R. and Valencia, A. (2005), 'Implementing the iHOP concept for navigation of biomedical literature', *Bioinformatics* Vol. 21, pp. ii252–ii258.
89. Eales, J., Pinney, J., Stevens, R. and Robertson, D. (2008), 'Methodology capture: Discriminating between the "best" and the rest of community practice', *BMC Bioinform*. Vol. 9, p. 359.
90. Steele, E., Tucker, A., 't Hoen, P. and Schuemie, M. (2009), 'Literature-based priors for gene regulatory networks', *Bioinformatics* Vol. 25, pp. 1768–1774.
91. MacCallum, R., Kelley, L. and Sternberg, M. (2000), 'SAWTED: Structure assignment with text description-enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons', *Bioinformatics* Vol. 16, pp. 125–129.
92. Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C.Y. *et al*. (2009), 'Identifying relationships among disease regions: Predicting genes at pathogenic SNP associations and rare deletions', *PLoS Genet*. Vol. 5, p. e1000534.
93. von Mering, C., Jensen, L., Snel, B. and Hooper, S. (2005), 'STRING: Known and prediction protein-protein associations, integrated and transferred across organisms', *Nucleic Acids Res*. Vol. 33, pp. D433–D437.