ELSEVIER

Status of text-mining techniques applied to biomedical text

Ramón A-A. Erhardt[1], Reinhard Schneider[2] and Christian Blaschke[1]

[1] Bioalma, Ronda de Poniente 4, 2° C-D, 28760 Tres Cantos, Madrid, Spain
[2] EMBL Heidelberg, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

Scientific progress is increasingly based on knowledge and information. Knowledge is now recognized as the driver of productivity and economic growth, leading to a new focus on the role of information in the decision-making process. Most scientific knowledge is registered in publications and other unstructured representations that make it difficult to use and to integrate the information with other sources (e.g. biological databases). Making a computer understand human language has proven to be a complex achievement, but there are techniques capable of detecting, distinguishing and extracting a limited number of different classes of facts. In the biomedical field, extracting information has specific problems: complex and ever-changing nomenclature (especially genes and proteins) and the limited representation of domain knowledge.

Effective knowledge management will be a key element for the success of the biotechnology and pharmaceutical industry in the years to come. Independent of the problem under study, revision and exploration of the knowledge already acquired is necessary for every researcher. The extensive use of high-throughput technologies, such as yeast-two hybrid-based methods, DNA expression arrays and mass spectrometry, generates massive amounts of data that in turn require efficient information retrieval before any analysis is attempted.

Scientific literature is a pivotal element in this knowledge management process because it is the largest and still the most reliable source of biomedical information. Technological advances and professional competition have contributed to the large volume of scientific articles, making it impossible for researchers to keep up with the literature.

Text mining (computer-executed automatic processing of large quantities of text) offers an interesting solution to that problem because it can reduce the time spent by researchers on reviewing the literature, by significantly covering many more scientific articles than those normally reviewed. This review will introduce some of the problems and challenges of text mining, and focus on those that are more frequently encountered in the application of these technologies.

The analysis of human language

The theoretical analysis of human language started long before the use of computers but has gained a lot of attention through research activities in computational linguistics (Box 1). Early natural language understanding systems like SHRDLU [1], dating back to the late 1960s, allowed interaction using English terms to manipulate a 3D world simulated by the system. SHRDLU not only understood instructions given by the user in its (restricted) vocabulary, but also possessed memory, could answer questions about the past and learn new facts (in its world). This was a powerful demonstration of artificial intelligence (AI), but led to excessive optimism that was soon lost when other systems attempted to deal with real-world ambiguity and complexity.

Natural language understanding has been referred to as an AI-complete problem, in analogy to NP (nondeterministic polynomial time)-completeness in complexity theory, which states that the difficulty of the computational problem is equivalent to designing a computer that is as intelligent as a human being [2].

It became evident over the past few decades that truly understanding human language requires extensive knowledge, not only about the language itself, but also about the surrounding world.

Corresponding author: Blaschke, C. (blaschke@bioalma.com)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.02.011

BOX 1

Glossary and definitions

Acronym. Abbreviations formed from the initial letter or letters of words, such as NATO (North Atlantic Treaty Organization) and HTML (Hyper Text Markup Language) with notable exceptions, such as XML, which stands for eXtensible Markup Language.
Anaphora. An anaphora is part of an expression referring to another part. For example, in 'the monkey took the banana and ate it'; 'it' refers to the banana.
Annotation. Extra information associated with a particular point in a document or other piece of information is called annotation. In the field of linguistics, annotations add information about the linguistic form of a particular text fragment. In this review, we use the term 'annotation' as reference to any information automatically added by a program – or manually by a domain expert – about specific named entities that appear in the text, or facts that were detected, such as interactions between proteins.
Computational linguistics. Computational linguistics is an interdisciplinary field dealing with the statistical and logical modelling of natural language from a computational perspective. Computational linguistics draws upon the involvement of linguists, computer scientists and professional experts in artificial intelligence, and cognitive psychologists and logicians, among others.
Co-occurrence. Co-occurrence is an event or situation that happens at the same time as, or in connection with, another. In the context of text analysis, two entities are said to co-occur if they appear together in the same piece of text, such as a sentence or a paragraph.
Disambiguation. In computational linguistics, word-sense disambiguation is the problem of determining in which sense a word with several distinct meanings is used in a given sentence. For example, consider the word 'bass', two distinct senses of which are a type of fish or tones of low frequency. And the sentences: 'The bass part of the song is very moving.' and 'I went fishing for some sea bass'.
Domain knowledge. In general, domain knowledge is the acquired comprehension valid and directly used for a preselected domain by a human, or an autonomous computer activity. An expert's domain knowledge is usually informal and ill-structured, and has to be transformed and encoded in computer programs and active data; for example, in a set of rules in knowledge bases, by knowledge engineers.
Linguistics. Broadly defined, linguistics is the scientific study of human language, and a linguist is someone who engages in this study.
Ontology. In information science, ontology is the product of an attempt to formulate an exhaustive and rigorous conceptual schema about a domain. This domain does not have to be the complete knowledge of that topic, but purely a field of interest decided upon by the creator of the ontology. Ontology is typically – but not necessarily – a hierarchical data structure containing all the relevant entities, and their relationships and rules within that domain. However, many existing cases of ontology use this definition with much less stringency, and are regarded to be shallow representations of the entities they define and the relations that exist between them, without guaranteeing complete consistency.
Parsing. Parsing is a process based on analyzing a sentence or string of symbols in a specific language, to determine its grammatical structure, with regards to a given formal grammar. A parser usually takes a lexical-analyzer-produced sequence of tokens as input, and afterwards builds a parse tree based on those tokens.
Precision. The term precision is used to mean the proportion of relevant documents from all results retrieved. 'How many of the things I consider to be true are actually true.' P = true positives ÷ (true positives + false positives).
Recall. Recall refers to the proportion of retrieved documents, out of all relevant results available. 'How much of the true things do I find.' R = true positives ÷ (true positives + false negatives).
Stemming. A stemmer is a program or algorithm that determines the morphological root of a given inflected – or sometimes derived – word-form. For example, a stemmer for English should identify the string 'cats' (and possibly 'catlike', 'catty', etc.) as based on the root 'cat', and the words 'stemmer', 'stemming', 'stemmed' as based on 'stem'.
Synonym. Synonyms are different words or phrases with similar or identical meanings.
Syntax. Syntax is the study of the rules, or 'patterned relations', that govern the way the words come together in a sentence. It relates to how different words (which are categorized as nouns, adjectives, verbs, etc.) are combined into clauses that, in turn, are joined to form sentences.
Token. A token is a primitive block of a structured text. Tokens are usually words – white spaces are usually ignored – but sometimes they are important and, therefore, tokenized. Tokenizing is frequently just the first step of interpreting text. Parsing follows tokenizing.

Nowadays, natural language processing (NLP) (Box 2) is a widely used term that does not imply a real understanding of language.

Finding the meaning of a sentence is complicated because of the ambiguity problem [3,4] – there are often several different possible meanings (e.g. in the phrase 'helicopter powered by human flies').

Many cases of ambiguity occur during speech recognition, syntactic and semantic analysis, and have to be resolved to provide a correct interpretation of an utterance. This ambiguity is largely resolved using contextual or general knowledge.

Lexical ambiguities

Lexical ambiguity is also called part of speech or category ambiguity. This problem arises because words can have more than one lexical class (e.g. 'bank' can be a verb or a noun).

Consider these two phrases [3]:

'Fying planes are dangerous.'

'Flying planes is dangerous.'

In the first sentence, 'flying' is an adjective (it is part of the noun phrase *flying planes*) and in the second, it is a verb. This ambiguity can be resolved by syntactical knowledge checking the subject-verb agreement. Similarly, consider 'bear left at zoo' – do you turn left when you get to the zoo, or did someone leave a bear there?

In a scientific context, consider:

(i) 'Somatosensory stimuli have complex timing relationships and are of long duration.'
(ii) 'In humans, telomeres are protected by shelterin, a complex of six proteins.'
(iii) 'Organic fractions extracted from digested sludge demonstrated a greater capacity to complex metals over mixed liquor extracts.'

These examples from biomedical journals show how the word 'complex' can be used as (i) an adjective, (ii) a noun and (iii) a verb.

Owing to lexical ambiguity, lexical tagging involves several automatic analyses and the recognition of grammatical context

BOX 2

Text mining overview:

Text mining, also known as intelligent text analysis, text-data mining or knowledge-discovery in text, generally refers to the extraction process of interesting and nontrivial information and knowledge, where an unstructured text is the source. The following is a (non-exhaustive) list of areas that belong or are related to text mining; the list is ordered alphabetically [93].

Document clustering and classification. When a set of documents is retrieved, these techniques organize the results into smaller groups, with the objective of assigning a specific document to one or more categories based on its contents. Document classification tasks can be divided into two types: supervised document classification, where some external mechanism (such as human feedback) provides information on the correct classification of documents, and unsupervised document classification, cases in which the classification must be carried out entirely without reference to external information.

Information extraction (IE). IE is a type of information retrieval whose goal is the automatic detection of assertions of restricted classes of facts, in a structured or semistructured form, from unstructured machine-readable documents. A typical application of IE is to scan a set of documents written in a natural language and to populate a database with the extracted information. Note that only restricted classes of facts are dealt with because IE does not claim to extract all the facts contained in a piece of text; it focuses on the retrieval of a predefined set of facts.

Information retrieval (IR). IR is a way of looking for information in documents, searching for documents themselves, for metadata that describe documents or for searches within databases, with the aim of finding text, sound, images or data. The task mainly consists of information retrieval from a larger set of documents that best matches a given query (in general determined as a list of words).

Name entity recognition (NER). NER is a subtask of information extraction that seeks to locate and classify single elements of the text into predefined categories, such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, and so forth.

Natural language processing (NLP). NLP is a subfield of artificial intelligence and linguistics. It studies the problems inherent to the processing, manipulation and understanding of natural language, and is devoted to making computers 'understand' statements previously written in human languages.

Question-answering (QA). QA is a type of IR and is sometimes regarded as the next step beyond current search engine technologies. Given a collection of documents, the system should be able to retrieve answers to (simple) questions posed in natural language.

Visualization. Visualization is not usually seen as part of text mining. But owing to the fact that – in most situations – users have to deal with large amounts of (text) data, and also the frequent complexity found in navigating through results, most systems have to provide new ways of visualizing the results, to allow the user to exploit them efficiently. Practical application of information visualization in computer programs involves selecting, transforming and representing abstract data in a form that enables human interaction for exploration and understanding. The interactivity and dynamics of the visual representation are important aspects of information visualization. Strong techniques enable the user to modify the visualization in real time, thus affording unparalleled perception of patterns and structural relations in the abstract data in question.

to check local constraints. Depending on systems, local constraint checking can be viewed as a part of lexical tagging because it contributes to selecting tags for words, as a part of syntactic parsing (Box 1) because it takes into account contextual constraints, or as an intermediate step [5].

Syntactic ambiguities

Syntactic (or structural) ambiguity is a property of sentences, which can be parsed in more that one way. For example:

'The students eat spaghetti with a fork.'

'The students eat spaghetti with ketchup.'

In the first sentence, 'with a fork' is attached to the students whereas in the second sentence, 'with ketchup' is attached to the spaghetti. This results in different interpretations and therefore structural differences in the parse tree that is used to represent the syntactic analysis of a sentence. This specific problem is called prepositional phrase (PP) attachment in the research literature (see the classic paper by Brill and Resnik [6]).

Syntactic ambiguity can be captured by returning several parse trees. In later processing steps, the correct tree has to be selected, applying semantic knowledge that would, for example, define spaghetti as food, fork as a tool and ketchup as a condiment. This knowledge has to be coded into software so that these sentences can be correctly analyzed on the syntactic level. The existing knowledge about similar objects like knives, ravioli or cheese would have to be encoded too, to analyze similar sentences.

In 'the cow was found by a stream by a farmer', the ambiguity is introduced by the different meanings of 'by', which in the first case refers to a physical location (the cow was located near a stream) and in the second case indicates how something is done (the farmer found the cow).

Here are two examples from biomedical documents:

'These data also suggest that AFB1 binds preferentially to DNA with an alternating G-C sequence compared to DNA with a sequence of contiguous Gs or Cs.'

'GMPPCP binds to tubulin with a low affinity relative to GTP or GDP.'

These examples show one possible case of syntactic ambiguity. In the first sentence, the prepositional phrase introduced by 'with' should be attached to the previous noun phrase ('DNA'). In the second sentence, the prepositional phrase should be attached to the noun phrase before the verb 'bind' ('CMPPCP'), and not to 'tubulin'. This ambiguity can only be resolved by knowing that G-C sequences are DNA characteristics and that low affinity is an attribute of a binding event.

Some advances have been made to learn this knowledge from the literature directly and to detect semantically related words by measuring the similarity of the surrounding contexts. These approaches could, for example, relate 'beer' and 'wine', because both of them appear in the context with words like 'drink', 'people', 'bottle' or 'make' [7–10]. Other approaches involve manual engineering of the knowledge domain, leading to more consistent and reliable models, but it are generally limited to very narrow domains or common English [11].

In the biomedical field, ontologies (Box 1) have been widely used for the structured organization of domain specific knowledge. A large number of medically oriented ontologies are summarized in the unified medical language system (UMLS [12]), maintained at

the US National Library of Medicine (www.ncbi.nlm.nih.gov). The UMLS is a controlled compendium of many vocabularies, providing a mapping structure between them. The 'Metathesaurus' forms the base of the UMLS and it is made up of over a million biomedical concepts and 5 million concept names, all of which are from over 100 controlled vocabularies and classification systems used in patient records, bibliographies, administrative health data and full-text databases. The Metathesaurus is organized by concept or meaning, and each concept has specific attributes that define the meaning. Identical or almost identical concepts are linked together with hierarchical context from the different vocabularies, and relationships between the concepts are explained and represented.

In biology, the gene ontology (GO, www.geneontology.org) has become very popular in recent years [13]. The GO is composed of three related ontologies covering basic areas of biological research: the molecular function of gene products, their role in biological processes and their physical structure as cellular components. Each ontology is constructed as a directed acyclic graph. The GO now contains over 17,000 terms referring to a wide variety of biological organisms. There is a significant body of literature on the development and use of the GO, and it has become a standard tool in the bioinformatics arsenal.

Despite these advances, we are far from computationally modelling human knowledge on a large enough scale for computers to 'understand' it.

Semantic ambiguities

Semantics covers the interpretation of the meaning of an utterance, as opposed to syntax, which describes the formal structure of an expression (Box 1). Semantic ambiguity refers to the problem of interpreting the meaning of a sentence and arises if words can have different meanings (e.g. 'bank' can refer to a river bank, a financial institution or the building where the financial institution is located [14]).

In these examples:

To seed/grow coffee;

To harvest/roast coffee;

To drink coffee.

'Coffee' can refer to a plant, its original grain or its derived commercial product. The verbs employed in these sentences modify the meaning of the object. This kind of ambiguity can only be resolved if the world knowledge is encoded into the system by specifying the semantic features for each item in the lexical index.

In 'I went to the bank to talk about a mortgage', knowledge about the meaning of 'mortgage' is needed to know that 'bank' means a financial institution and not a river bank.

In 'The normal, scrapie, and CJD isoforms of scrapie-associated proteins share common epitopes with varying degrees of interspecies homology', CJD is a protein.

However, in 'In addition, we examined normal human brain and brain tissues from patients with CJD, kuru, Alzheimer's disease, and idiopathic chronic encephalitis', CJD means Creutzfeldt–Jacob disease. Knowledge about the context of the sentences (i.e. the words 'isoforms' in the first and 'patient' and 'disease' in the second) can be used to resolve these ambiguities.

Semantic ambiguities are best handled using a 'vague' representation language for sentence meaning. A computer parser could solve these ambiguities by specifying the semantic features of each item in the lexical entry (for a general overview see [3,15]).

Implications of ambiguity in language

Because of the ambiguities in natural languages, an immense amount of knowledge needs to be encoded in computer readable forms, thus letting computers 'understand' and correctly interpret human language. As mentioned before, this has not been achieved beyond fairly restricted systems. Therefore, NLP systems have to come up with approximations and are restricted to specific areas where limited domain knowledge (Box 1) can be built into the system.

Extracting information from text

As already pointed out, the limitations of natural language understanding are tied to the knowledge that can be encoded into a computer. Representing the entire knowledge available to a human in computer readable form is, at least at this time, not realistic. But the extraction of a restricted set of facts from a limited knowledge domain is possible today, and advances in different directions have been made.

Named entity recognition

Named entity recognition (NER) is a prerequisite of information extraction (Box 2) and aims to locate and classify tokens (Box 1) in text into predefined categories (e.g. the names of people, organizations and locations, and expressions of times, quantities, monetary values, percentages, etc.). In the biomedical sector, the majority of existing works have been focused on the detection of genes and proteins and less attention has been paid to chemical compounds, diseases and general biomedical terms. The ambiguity of language itself and the constant changes and advances in the biomedical nomenclature leave us with this difficult task that, nevertheless, is the basis for a successful extraction of information.

Simplification of the NLP problem: information extraction

It can be postulated that we will not be able to mimic the knowledge of a typical human being with a computer – at least not in the near future. A more realistic goal is the extraction of specific information in a well-defined and restricted domain. This is done by information extraction systems.

Information extraction (IE) is a technology based on the analysis of natural language used to extract clearly defined pieces of information. The systems often apply rules, patterns, frames or templates (these are just synonymous terms used by different authors but they basically refer to the same principle) to locate the information in the text. The produced structured output is either directly displayed for the user or stored in a database for further analysis. Typically, only parts of a text are relevant, and out of a relevant sentence often only a part contains the significant information. Most IE systems analyze texts in a sequential process with increasing complexity, which usually includes the following steps: lexical and morphological processing; recognition and classification of named entities; parsing of larger syntactic constituents; resolution of anaphora (Box 1); and the ultimate extraction of domain-relevant events and relationships from the text.

The core of an IE system consists of rules that are used to extract specific information. There are two basic approaches to the design

of these rules. They can be classified into a knowledge engineering approach and an automatic training approach. The knowledge engineering approach is characterized by the manual development of these rules by a knowledge engineer. This role is often assumed by linguists, who know how to express the rules, complemented either with the help of a domain expert or with an annotated corpus of domain-relevant texts. An automatically trained system requires a relatively large corpus of training text that is already annotated for the wanted knowledge domain. The respective rules are learned from these examples (for a good introduction to IE see [16]).

Traditionally biomedical IE systems are based on knowledge engineering but in recent years we have seen some advances in acquiring them automatically. Language is not only ambiguous but also very flexible, and therefore it is possible to express the same (or very similar) facts in very different ways.

Consider the following examples of protein–protein interaction:

Protein X is activated by protein Y;
Protein X is phosphorylated by protein Y;
Protein Y phosphorylates protein X;
Protein Y regulates protein X;
Under Z conditions we observed an increased affinity of protein X to protein Y;
Protein X/protein Y activation was reported.

This is only a small subset of ways of describing protein interactions, and we would already need a handful of rules to cover this example. The more linguistic parameters are considered (e.g. the part of speech of the words), the more rules are necessary to extract the desired information and this task quickly becomes overwhelming.

Developing the rules manually is an intensive and difficult process because knowledge about the IE system and the domain is needed. Automatic systems generally suffer from sparse training data, because many examples are needed if the system is expected to generate reliable rules. Furthermore, these systems are limited in scope because each new question needs to go through the steps of engineering the rules or annotating training corpora again. The limited set of rules available to an IE system typically covers only the most frequent appearances and can not extract large amount of information if the rules do not fit.

Text mining applications for biomedical research

Text mining was first applied to the biological research area in the late 1990s. Owing to the complexity of the biomedical nomenclature and the importance of detecting genes and proteins as a basic building block for information extraction systems, this task has awakened considerable interest. Although many biologically interesting applications exist, most information extraction systems have focused on the detection of protein–protein interactions, and less work has been done in other fields.

Biomedical named entity recognition

A prerequisite for indexing and retrieving relevant documents and information from the literature is the successful identification of biomedical terms [called 'named entity recognition' (NER) applied to the specific domain at hand]. Based on the complexities of a dynamically changing biomedical terminology, term identification has been recognized as the current bottleneck in text mining. Consequently term identification has become an important research topic in natural language processing.

Dictionaries containing the vocabulary used in the domain are an important source for NER. When combining information from EntrezGene and UniProt [17] 32,777 genes are found with a total of 168,952 different names in the human genome alone. This would result in 5.2 synonyms per gene (Box 1). Of these, 4,930 names are ambiguous because they refer to more than one gene.

Errors in these steps can lead to problems in later processes. Jennsen *et al.* [18] reported that ~40% of the errors in extracted gene networks resulted from incorrectly detected gene names, due to abbreviations or words misinterpreted as gene symbols. The main reason (85% of cases) for missed interactions was that genes were overlooked by the system.

Processes involved in the detection of biomedical entities

Named entity (NE) detection and classification tries to find items of interest in a text and to classify them into predefined categories, such as gene, protein or chemical compound, in the biomedical domain. In these cases, the challenge lies in the determination of start and end boundaries of the names and assignment to the correct class.

Disambiguation requires the determination of the specific identity of the detected entities and the mapping to the unique concept to which they refer (e.g. unique identifiers of a database like EntrezGene) (Box 1). This process is also known as term normalization. Because of the ambiguities in the biomedical nomenclature, the system needs to distinguish between different meanings of the names.

These problems are obviously related, seeing that one can not address the disambiguation without having first detected the entities in the text. The detection of gene and protein names was one of the first text mining tasks to be addressed in the biomedical field [19–22]. The detection of other biological entities, like chemical substances [23] or diseases [24–27], has attracted much less attention. There was also much less variety because most systems [28–30] were developed using the framework of the GENIA project [31] from the University of Tokyo, or based on the UMLS.

Human language in general – and biomedical nomenclature in particular – show ambiguities, and the detected entities have to be normalized and associated with a specific biological object. For example, the name 'alcohol dehydrogenase' is a valid synonym for 111 different fly genes and the gene symbol *PAP* can refer to the following different human genes (symbols and names, according to EntrezGene): *PAPOLA* or poly(A) polymerase α; *MRPS30* or mitochondrial ribosomal protein S30; *REG3A* or regenerating islet-derived 3 α (also know as pancreatitis-associated protein); *PDAP1* or PDGFA associated protein 1; *TUSC2* or tumor suppressor candidate 2; *DDEF1* or development and differentiation enhancing factor 1; *DDEF2* or development and differentiation enhancing factor 2; and *ACPP* or acid phosphatase, prostate.

A text token like PAP needs to be disambiguated before the system can proceed with any higher-level analysis steps [32–34]. Table 1 shows some occurrences of PAP citing different genes.

TABLE 1

Searches on the term 'PAP' citing different genes

Gene	Sentence	Pubmed ID
ACPP	Purified human prostate acid phosphatase (**PAP**) was used to generate a specific monoclonal antibody (FC 3001) for detection of PAP expressed by some prostatic carcinomas.	8305218
REG3A	We have shown that HIP is in fact the pancreatitis-associated protein (**PAP**) and provided a phylogenetic analysis of the free CRD lectins.	8076648
PAPOLA	We also show that 160K binds specifically to both the 77K (suppressor of forked) subunit of the cleavage factor CstF and to poly(A) polymerase (**PAP**).	7590244
PDAP1	The protein copurified with platelet-derived growth factor (PDGF)-A and was therefore termed PDGF-associated protein (**PAP**).	8780057

Because of the challenges of biomedical entity disambiguation, very little work has been done to address this problem. Applications not related to gene names have mostly focused on disambiguating general medical terms in the UMLS [35–37]. Gene name disambiguation has gained considerable interest over the last years and several interesting solutions have been proposed [29,30,38–41].

Acronym resolution

An additional problem of the biomedical literature is the abundance of acronyms (Box 1) and abbreviations. Because of their limited length, these acronyms are very often identical to gene symbols, and increase the already existing ambiguity of the gene nomenclature. A system not only has to distinguish different genes with identical names, but also has to detect if a likely gene name refers to a completely different entity. There is evidence of cell lines and viruses often having identical names as genes [42], and 80% of the abbreviations defined in the UMLS have ambiguous occurrences in Medline [43].

SCT, for example, is the official gene symbol for the human gene secretin. On the other hand, most documents in Medline citing the item 'SCT' can openly refer to more than 100 different meanings like 'stem cell transplant', 'salmon calcitonin' or 'stair climbing test' (see the Biomedical Acronym Resolver, http://invention.swmed.edu/argh). The reason for this is the introduction of acronyms to abbreviate complex terms without checking if these acronyms overlap with the gene nomenclature. For a human reader this would be no problem, but an automatic system has to apply specific methods to resolve this dilemma.

Several approaches have been proposed to automatically extract acronyms from the literature [44–47]. We found very few cases where efforts in acronym extraction and gene name disambiguation have been joined [48], and we expect that the performance of gene name disambiguation will improve once these two approaches are integrated.

System performance

There has been an increasing amount of work on text mining from the literature, but it is currently difficult to compare the systems because they use different datasets and perform different tasks. In the past, evaluations have been very useful to evaluate text-mining systems under controlled conditions and to monitor progress over time (www.itl.nist.gov/iaui/894.02/related_projects/muc). The BioCreative (critical assessment of information extraction in biology) assessment [49–51] provided a systematic evaluation for a set of biological text mining tasks. We will restrict the discussion to the job that covered the detection of gene and protein names (named entity detection and classification [52–54]) and the disambiguation problem of gene and protein names (disambiguation or normalization [55–58]).

The authors describe several typical problems that the participants encountered, such as tokenization errors (definition of word boundaries), inconsistent and noisy training data and complex gene nomenclatures with extensive ambiguity, overlap of gene names with English terms, and complex multiword names. Nevertheless, named entity detection and classification systems achieved F measures of over 0.80. The F measure is the harmonic mean of recall and precision (Box 1) and the achieved value is only somewhat lower than figures known from the news wire domain. The result of the disambiguation task was heavily dependent on the analyzed organism. The results were very good for yeast and of poorer quality for fly and mouse. This can be explained, at least in part, by the simpler yeast nomenclature and the fact that authors seem to respect the established standard. The overall impression is that there is still some room for improvement, but tools for automated gene name identification and normalization are becoming ready to be incorporated as building blocks in other text mining systems.

Similarly, the BioCreative assessment in the Text REtrieval Conference (TREC; http://trec.nist.gov/) is an ongoing series of workshops focusing on a list of different information retrieval (IR, see Box 2) research areas, or 'tracks'. The goal of the genomics track is to study the retrieval of genomic data, not just gene sequences but also supporting documentation, such as research papers and laboratory reports [59].

Extraction of biomedical information

Because of the challenges arising from building and maintaining IE systems, these actions have been applied to a limited set of facts, and the most notable successes are the extraction of protein–protein interactions and gene regulation events. The detection of protein interactions in Medline abstracts was the first applications of IE technologies to in this field [60]. The early systems used regular expression-based patterns [60–62], whereas later systems applied sentence parsers and several different grammar formalisms [63–69].

Much less attention has been paid to the extraction of other facts. It is difficult to define clear boundaries between the different systems, but some patterns emerge. The following applications are published so far: extraction of information about functional

aspects of genes and proteins [70–72], an algorithm to detect relationships to drugs [21], and several medical applications [73–76].

In the past few years, systems that implement automatic training have found their way into the biological domain but applications have so far been limited to protein–protein interactions [77]. During the recent LLL05 (Learning Language in Logic) challenge, some interesting results were shown [78].

The LLL05 task was to learn rules to extract protein–gene interactions in the form of relations from biological text, using machine-learning methods. The best systems yielded precision and recall values of around 50%, which is relatively high compared with similar challenges on event or relation extraction. But the test data were probably carefully selected with the aim of keeping the underlying biological models simple, and richly annotated training data by domain experts is needed for these methods to work satisfactorily.

Co-occurrence based text mining

The ambiguities of natural language and their great flexibility leads NLP systems to focus on very specific problems in the biomedical domain, namely protein interactions as well as a small number of other applications. Instead of trying to solve the rather challenging problem of understanding the meaning of specific sentences, co-occurrence-based NLP methods (Box 1) propose a radically different approach. The information extraction is done by looking at word or term distributions, or by analyzing of the information content of different passages.

One can, for example, observe the co-occurrence of biological entities in sentences or paragraphs and calculate how unlikely it is to observe a certain level of co-occurrences by chance. The more unlikely the observed event, the stronger the relation between the entities is valued by the system. Using this approximation, gene association networks can be created, not specifying the precise relationships between the genes but organizing the lit-

TABLE 2

List of the first 25 words extracted from the documents where the gene APOE was detected[a]

	Stem of word	Words belonging to the stem	Frequency	Relevance score
1	apo	apoes, apoed, apo, apoe, apos	0.77	1079.8
2	apolipoprotein	apolipoproteine, apolipoprotein, apolipoproteins	0.86	1079.2
3	epsilon4	epsilon4	0.16	836.9
4	apoe4	apoe4	0.09	637.6
5	e4	e4	0.11	409.9
6	e	e	0.81	379.0
7	alzheimer's	alzheimer's	0.35	332.1
8	genotyp	genotyp, genotyps, genotype, genotyper, genotypes, genotyped, genotypic, genotypers, genotyping, genotypized, genotypable, genotypings, genotypical, genotypeable, genotypically, genotypability, genotypization	0.40	294.6
9	allel	allel, allels, allele, alleled, allelic, allells, allelle, alleles, allellic, alleleic, allelles, allelism, allelisms, allellism, allelically	0.45	291.2
10	ad	ad	0.27	268.6
11	lipoprotein	lipoproteins, lipoproteine, lipoprotein, lipoproteinic	0.32	246.7
12	epsilon	epsilones, epsilons, epsilon, epsilonal, epsilone	0.13	208.4
13	amyloid	amyloid, amyloids, amyloide, amyloidal, amyloidic, amyloiditis, amyloidation, amyloidization	0.15	187.8
14	alzheim	alzheimic, alzheimer, alzheimers	0.09	174.2
15	ldl	ldl, ldls	0.16	169.0
16	vldl	vldls, vldl	0.07	169.0
17	cholesterol	cholesterol, cholesteroll, cholesterols, cholesterole, cholesterolic	0.27	166.7
18	polymorph	polymorph, polymorphe, polymorphs, polymorphes, polymorphic, polymorphous, polymorphism, polymorphics, polymorphisms, polymorphously, polymorphically	0.27	164.2
19	hdl	hdls, hdl	0.12	143.6
20	dementia	dementiae, dementia, dementias	0.13	135.5
21	triglycerid	triglyceridic, triglycerid, triglycerides, triglyceride, triglycerids	0.13	125.9
22	lipid	lipid, lipids, lipide, lipides, lipidic, lipidous, lipidate, lipidates, lipidated, lipidized, lipidizing, lipidation, lipidating, lipidization	0.24	107.5
23	densiti	densitys, densities, density	0.23	99.5
24	atherosclerosi	atherosclerosis	0.09	82.0
25	plaqu	plaquing, plaques, plaque, plaqued	0.08	74.3

[a] At the time of analysis, there were 5181 documents for this gene in our database. The table shows the root of the term (produced by stemming), the list of words it corresponds to, the frequency of the word in the document set (fraction of the 5181 documents where the word appears) and the relevance score calculated by the system.

erature in a way that makes exploration a lot easier [18,79]. In a similar way, keywords can be extracted from a set of documents that guide the user and are the basis for a scoring scheme, which allows the interactive sorting of sentences and documents by the user. [80]

Linguistic analysis is too narrow for applications with a broad scope, so statistical methods can produce very useful results. In the biomedical domain, these methods have been used in problems like the data analysis of DNA arrays [81–84], gene clustering based on function similarity [85,86], extraction of functional information for genes and proteins [87–90], improving remote homology searching [91] and prediction of genes related to certain diseases [92].

Table 2 shows the typical results that systems based on statistical methods can achieve. The first 25 of the extracted keywords for the gene *ACHE* (an Acetylcholinesterase that is known to be implicated in disease processes related to neurodegenerative diseases) are listed. The keywords serve as a summary of the themes covered in relation with the gene *ACHE* and are the basis for the scoring of sentences and documents (Table 3).

The visualization (Box 2) of co-occurrence networks illustrates complex relationships between large numbers of elements. Figure 1 shows how two genes are related to Alzheimer's disease (the Achetylcholinesterase, *ACHE*, and the Apolipoprotein E, *APOE*) based on their associations with chemical substances. This type of visualization shows the 'big picture' that leads to an easy characterization of the relations before going into more details.

Computational linguistics has shown to be extremely valuable for the detection of protein interaction or gene regulation events, however it's limitation to specific questions and its intensive

TABLE 3

The ten highest ranked documents and sentences for the human gene APOE based on the selection of cholesterol and atherosclerosis from the extracted keywords

Rank	Titles	Relevance score
1	Rapid regression of atherosclerosis induced by liver-directed gene transfer of ApoE in ApoE-deficient mice.	4019.6
2	Reduction of isoprostanes and regression of advanced atherosclerosis by apolipoprotein E.	3887.6
3	Acute regression of advanced and retardation of early aortic atheroma in immunocompetent apolipoprotein-E (apoE) deficient mice by administration of a second generation [E1(−), E3(−), polymerase(−)] adenovirus vector expressing human apoE.	3628.2
4	Macrophage-specific expression of human apolipoprotein E reduces atherosclerosis in hypercholesterolemic apolipoprotein E-null mice.	3501.3
5	Hepatic expression of apolipoprotein E inhibits progression of atherosclerosis without reducing cholesterol levels in LDL receptor-deficient mice.	3454.9
6	Blockade of endothelin receptors reduces diet-induced hypercholesterolemia and atherosclerosis in apolipoprotein E-deficient mice.	3341.8
7	Liver-directed gene transfer and prolonged expression of three major human ApoE isoforms in ApoE-deficient mice.	3328.2
8	Isoform-specific effects of apolipoprotein E on atherogenesis: gene transduction studies in mice.	3260.9
9	Overexpressed lipoprotein lipase protects against atherosclerosis in apolipoprotein E knockout mice.	3200.9
10	The independent correlation of the impact of lipoprotein(a) levels and apolipoprotein E polymorphism on carotid artery intima thickness.	3188.9
Rank	**Sentences**	**Relevance score**
1	In addition, although the combination of cholesterol reduction and apoE expression significantly reduced atherosclerosis, its effects were no greater than with expression of the LDLR or apoE alone.	7096.4
2	Apolipoprotein E (apoE) polymorphism affects plasma cholesterol and may influence risk of atherosclerosis.	3940.3
3	Apolipoprotein E (apoE) is a major constituent of plasma lipoprotein that functions in lipid transport and redistribution (reverse cholesterol transport) and probably plays an important role in inhibiting the development and/or progression of atherosclerosis.	3765.0
4	Apolipoprotein E (apoE) reduces mouse atherosclerosis progression independent of plasma cholesterol level effects.	3591.3
5	Variation at the apolipoprotein E (apo E) gene locus affects cholesterol concentrations, the risk for atherosclerosis and Alzheimer disease (AD), and is associated with longevity in Caucasians.	3567.5
6	Effect of macrophage-derived mouse ApoE, human ApoE3-Leiden, and human ApoE2 (Arg158–>Cys) on cholesterol levels and atherosclerosis in ApoE-deficient mice.	3456.3
7	Cholesterol efflux from macrophages to apoE has been shown to decrease foam cell formation and prevent atherosclerosis.	3124.7
8	The effect of monocyte/macrophage-derived wild-type mouse apolipoprotein E (apoE), human apoE3-Leiden, and human apoE2 on serum cholesterol levels and the development of atherosclerosis in apoE-deficient (apoe−/−) mice was investigated by using bone marrow transplantation (BMT).	2909.1
9	The inverse relationship of E2E3 with carotid artery atherosclerosis seems to be independent of serum apoE and total and HDL cholesterol levels.	2665.6
10	A small number of proteins have been shown in vitro to be upregulated by cellular cholesterol loading, including apolipoprotein E (apoE) and the recently cloned HDL-binding protein (HBP), but only apoE has been shown to be upregulated in cholesterol-loaded cells in atherosclerosis.	2602.0

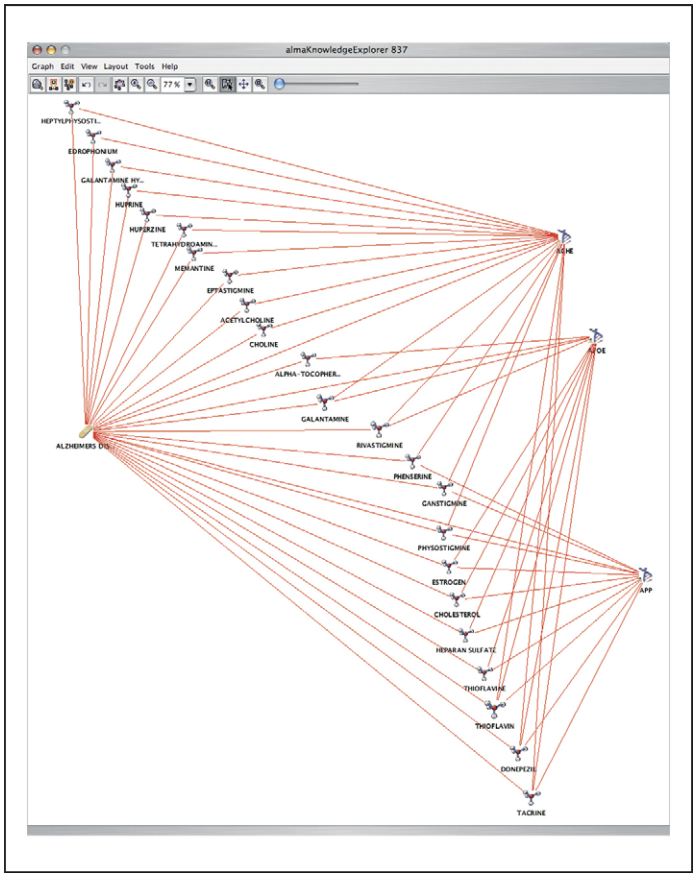

FIGURE 1

Co-occurrence map that shows a direct relation of Alzheimer's disease and ACHE, APOE and APP genes. The map also shows the relation of the disease through chemical compounds. The compound–gene relationship gives an idea of the compound mechanisms. From the sentences that support this graph (data not shown) we can see that a group of chemicals related to *ACHE* are inhibitors of this enzyme and are suitable treatments for Alzheimer's disease. Many of those, such as donepezil or tacrine, are also related to *APOE*. Different genotypes of this gene might affect drug activity. The mechanism of this modification appears to be related with the *APP* gene.

customization requirements make the co-occurrence analysis today's most effective information retrieval solution.

Conclusion and outlook

Keeping up with the literature and understanding what has been published is crucial for scientific advancement. However, investigators often have to focus on research and have limited time to spend on literature searches. Although we have argued that computers will not, at least not in the near future, be able to truly understand natural language, text-mining systems have considerably improved. A variety of systems have become available for specific domains that tap into the wealth of information in the literature and help researchers keep up with managing the vast amount of information that is available today.

Some of the key issues pointed out are:

(i) Literature mining is domain-specific for two reasons. First, the objects described are specific for that domain (proteins and diseases in the biomedical domain). Second, the relationships that exist between these entities also tend to be very specific (protein interactions).

(ii) Detection of biological entities. Owing to the complexity of the biomedical nomenclature and its ever-evolving character, much care has to be taken at this step. Acronyms often constitute valid gene names and many genes have the same name, making a correct detection in the text difficult. Text-mining addresses this problem and users should take care to choose systems offering state-of-the-art technology that goes beyond dictionary- or vocabulary-based approaches. The disambiguation and correct classification of the relevant entities appearing in the documents is clearly essential for any information to be extracted later on.

(iii) Computational linguistics versus co-occurrence-based text-mining methods. Because of the complexities of languages, linguistic approaches are limited to restricted domains and customization to a new set of facts is a very labour-intensive task. They are a good choice for the domains that they have been developed for (e.g. protein interactions or genetic interactions) because they provide very precise results and no user intervention is required. However, they are impossible to implement when the questions are not defined beforehand. In this case, statistical methods based on co-occurrences and other measures can be very fast and useful for getting precise information. An additional advantage is that the recovery rates are very high and there is a low risk of missing relevant information.

Acknowledgements

This work was supported in part by the EU projects Alvis IST-1–002068-STP and GeneFun LSHG-CT-2004–503567.

References

1 Terry, W. (1971) Procedures as a representation for data in a computer program for understanding natural language. *MIT AI Technical Report* 235

2 Mallery, J.C.M. (1988) Thinking about foreign policy: finding an appropriate role for artificially intelligent computers *The Annual Meeting of the International Studies Association* St. Louis, MO, USA

3 Allen, J. (1995) *Natural Language Understanding* ((2nd edn)), Benjamin/Cummins Publishing

4 Russell, S. and Norvig, P. (1995) *Artificial Intelligence: A Modern Approach*. Prentice Hall

5 Gorfein, D.S., ed. (2001) *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*, APA Books

6 Eric Brill and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th conference on Computational Linguistics*, 1198–1204

7 Pantel, P. and Lin, D. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In Proc. of ACL-2000, Hong Kong

8 Dagan, I. *et al.* (1995) Contextual word similarity and estimation from sparse data. *Computer, Speech and Language* 9, 123–152

9 Li, H. and Abe, N. (1998) Word clustering and disambiguation based on co-occurrence data. In Proceedings of COLING

10 Lin, D. (1998) Automatic retrieval and clustering of similar words. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistic* 2, 768–774

11 Fellbaum, C. (1999) *WordNet: an Electronic Lexical Database*. MIT Press
12 Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32 (Database issue), D267–D270
13 Bada, M. *et al.* (2004) A short study on the success of the Gene Ontology. *J. Web Semantics* 1, 235–240
14 Ginter, F. *et al.* (2004) New Techniques for Disambiguation in Natural Language and Their Application to Biological Text. *J. Mach. Learn. Res.* 5, 605–621
15 Jurafsky, D. and Martin, J.H. (2000) *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall
16 Appelt, D.E. and Israel, D.J. (1999) Introduction to information extraction technology. *In Proceedings of the 16th International joint Conference on Artificial Intelligence*, Stockholm, Sweden
17 Jensen LJ, Saric J, Bork P. (2003) Utilizing literature for biological discovery. *E-BioSci/ORIEL Annual Workshop*. Varenna, Italy
18 Jenssen, T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28
19 Fukuda K. *et al.* (1998) Toward Information Extraction: Indentifying Protein Names from Biological Papers. *In Proceedings of the Pacific Symposium on Biocomputing*. 3, 707–718.
20 Proux, D. *et al.* (1998) Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Inform. Ser. Workshop Genome Inform.* 9, 72–80
21 Rindflesch, T.C. *et al.* (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* 5, 517–528
22 Tanabe, L. and Wilbur, W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics* 18, 1124–1132
23 Wilbur, W.J. *et al.* (1999) Analysis of biomedical text for chemical names: a comparison of three methods. *Proc. AMIA Symp.* 176–180
24 Adamic, A.L. *et al.* (2002) A Literature Based Method for Identifying Gene-Disease Connections. *Proceedings of the IEEE Comput. Soc. Bioinform. Conf.* 1, 109–117
25 Grishman, R. *et al.* (2002) Information extraction for enhanced access to disease outbreak reports. *J. Biomed. Inform.* 35, 236–246
26 Friedman, C. *et al.* (2004) Automated encoding of clinical documents based on natural language processing. *J. Am. Med. Inform. Assoc.* 11, 392–402
27 Wren, J.D. and Garner, H.R. (2004) Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics* 20, 191–198
28 Torii, M. *et al.* (2004) Using name-internal and contextual features to classify biological terms. *J. Biomed. Inform.* 37, 498–511
29 Zhou, G. *et al.* (2004) Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics* 20, 1178–1190
30 Lee, K.J. *et al.* (2004) Biomedical named entity recognition using two-phase model based on SVMs. *J. Biomed. Inform.* 37, 436–447
31 Kim, J.D. *et al.* (2003) GENIA corpus–semantically annotated corpus for bio-textmining. *Bioinformatics* 19 (Suppl 1), i180–i182
32 Mons, B. (2005) Which gene did you mean? *BMC Bioinformatics* 6, 142
33 Chen, L. *et al.* (2004) Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics* 21, 248–256
34 Tuason, O. *et al.* (2004) Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pac. Symp. Biocomput.* 8, 238–249
35 Schuemie, M.J. *et al.* (2005) Word sense disambiguation in the biomedical domain: an overview. *J. Comput. Biol.* 12, 554–565
36 Leroy, G. and Rindflesch, T.C. (2005) Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *Int. J. Med. Inform.* 74, 573–585
37 Liu, H. *et al.* (2001) Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *J. Biomed. Inform.* 34, 249–261
38 Cohen, A. (2005) Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. *Proceedings of the Joint ACL Workshop and BioLINK SIG (ISMB) on Linking Biological Literature Ontologies and Databases* Detroit, MI, USA
39 Schijvenaars, B.J. *et al.* (2005) Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics* 6, 149
40 Morgan, A.A. *et al.* (2004) Gene name identification and normalization using a model organism database. *J. Biomed. Inform.* 37, 396–410
41 Podowski, R.M. et al. (2004) AZuRE, a Scalable System for Automated Term Disambiguation of Gene and Protein Names. *In 3rd International IEEE Computer Society Computational Systems Bioinformatics Conference. IEEE Computer Society*, 415–424
42 Tanabe, L. and Wilbur, W.J. (2002) Tagging gene and protein names in full text articles. *Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain*. 9–13
43 Liu, H. *et al.* (2002) A study of abbreviations in MEDLINE abstracts. *Proc. AMIA Symp.* 464–468
44 Wren, D.J. *et al.* (2005) Biomedical Term Mapping Databases. *Nucleic Acids Research (Database Issue)* 33, 289–293
45 Chang, J.T. *et al.* (2002) Creating an online dictionary of abbreviations from MEDLINE. *J. Am. Med. Inform. Assoc.* 9, 612–620
46 Yu, H. *et al.* (2002) Mapping abbreviations to full forms in biomedical articles. *J. Am. Med. Inform. Assoc.* 9, 262–272
47 Pustejovsky, J. *et al.* (2001) Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo* 10, 371–375
48 Podowski, R.M. *et al.* (2005) Suregene, a scalable system for automated term disambiguation of gene and protein names. *J Bioinform Comput Biol.* 3, 743–770
49 Hirschman, L. *et al.* (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics* 6 (Suppl 1), S11
50 Hirschman, L. *et al.* (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 6 (Suppl 1), S1
51 Yeh, A. *et al.* (2005) BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics* 6 (Suppl 1), S2
52 Kinoshita, S. *et al.* (2005) BioCreAtIvE Task1A: entity identification with a stochastic tagger. *BMC Bioinformatics* 6 (Suppl. 1), S4
53 Finkel, J. *et al.* (2005) Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics* 6 (Suppl. 1), S5
54 Hakenberg, J. *et al.* (2005) Systematic feature evaluation for gene name recognition. *BMC Bioinformatics* 6 (Suppl. 1), S9
55 Crim, J. *et al.* (2005) Automatically annotating documents with normalized gene lists. *BMC Bioinformatics* 6 (Suppl. 1), S13
56 Hanisch, D. *et al.* (2005) ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics* 6 (Suppl. 1), S14
57 Fundel, K. *et al.* (2005) A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics* 6 (Suppl. 1), S15
58 Tamames, J. (2005) Text Detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics* 6 (Suppl. 1), S10
59 Hersh, W. *et al.* (2004) Enhancing access to the bibliome: the TREC genomics track. *Medinfo* 11, 773–777
60 Blaschke, C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol.* 30A, 60–67
61 Ng, S. and Wong, M. (1999) Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics* 10, 104–112
62 Wong, L. (2001) PIES, a protein interaction extraction system. *Pac. Symp. Biocomput.* 6, 520–531
63 Daraselia, N. *et al.* (2004) Extracting human protein interactions from MEDLINE using a fullsentence parser. *Bioinformatics* 22, 604–611
64 Friedman, C. *et al.* (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformaticas* 17 (Suppl. 1), S74–S82
65 Rindflesch, T.C. *et al.* (1999) Mining molecular binding terminology from biomedical text. *Proc. AMIA Symp.* 34, 127–131
66 Saric, J. *et al.* (2005) Extraction of regulatory gene/protein networks from Medline. *Bioinformatics July* 26
67 Thomas, J. *et al.* (2000) Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.* 541–552
68 Wattarujeekrit, T. *et al.* (2004) PASBiol.: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics* 5, 155
69 Chiang, J.H. *et al.* (2004) GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics* 20, 120–121
70 Adamic, L.A. *et al.* (2002) A literature based method for identifying gene-disease connections. *Proc. IEEE Comput. Soc. Bioinform. Conf.* 1, 109–117
71 Divoli, A. and Attwood, T.K. (2005) BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics* 21, 2138–2139
72 Koike, A. *et al.* (2004) Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics* 21, 1227–1236
73 Hahn, U. *et al.* (2002) MEDSYNDIKATE–a natural language system for the extraction of medical information from findings reports. *Int. J. Med. Inform.* 67, 63–74
74 do Amaral, M.B. *et al.* (2000) NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. *Proc. AMIA Symp.* 76–80
75 Friedman, C. *et al.* (1999) Representing information in patient reports using natural language processing and the extensible markup language. *J. Am. Med. Inform. Assoc.* 6, 76–87

76 Libbus, B. and Rindflesch, T.C. (2002) NLP-based information extraction for managing the molecular biology literature. *Proc. AMIA Symp.* 445–449
77 Donaldson, I. *et al.* (2003) PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* 4, 11
78 Nédellec, C. (2005) Learning language in logic - genic interaction extraction challenge, (Cussens, J. and Nédellec, C., eds) *In Proceedings of the 4th Learning Language in Logic Workshop (LLL05)* 31-37, Bonn, Germany
79 Stapley, B.J. and Benoit, G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.* 5, 529–540
80 Andrade, M.A. and Valencia, A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14, 600–607
81 Shatkay, H. *et al.* (2000) Genes, themes, and microarrays: using information retrieval for largescale gene analysis. *Intell. Syst. Mol. Biol.* 8, 317–328
82 Blaschke, C. *et al.* (2001) Mining functional information associated with expression arrays. *Funct. Integr. Genomics* 1, 256–268
83 Raychaudhuri, S. *et al.* (2003) The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res.* 31, 4553–4560
84 Rubinstein, R. and Simon, I. (2005) MILANO–custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics* 6, 12
85 Blaschke, C. and Valencia, A. (2002) Automatic ontology construction from the literature. *Genome Informatics* 13, 201–213
86 Homayouni, R. *et al.* (2005) Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics* 21, 104–115
87 Raychaudhuri, S. *et al.* (2002) Using text analysis to identify functionally coherent gene groups. *Genome Res.* 12, 1582–1590
88 Tamames, J. *et al.* (1998) EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics* 14, 542–543
89 Perez, A.J. *et al.* (2004) Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics* 20, 2084–2091
90 Raychaudhuri, S. *et al.* (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.* 12, 203–214
91 Chang, J.T. *et al.* (2001) Including biological literature improves homology search. *Pac. Symp. Biocomput.* 6, 374–383
92 Perez-Iratxeta, C. *et al.* (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* 31, 316–319
93 de Bruijn, B. and Martin, J. (2002) Getting to the (c)ore of knowledge: mining biomedical literature. *Int. J. Med. Inform.* 67, 7–18