DRUG DISCOVERY TODAY BIOSILICO

Text-mining approaches in molecular biology and biomedicine

Martin Krallinger, Ramon Alonso-Allende Erhardt and Alfonso Valencia

Biomedical articles provide functional descriptions of bioentities such as chemical compounds and proteins. To extract relevant information using automatic techniques, text-mining and information-extraction approaches have been developed. These technologies have a key role in integrating biomedical information through analysis of scientific literature. In this article, important applications such as the identification of biologically relevant entities in free text and the construction of literature-based networks of protein–protein interactions will be introduced. Also, the use of text mining to aid the interpretation of microarray data and the analysis of pathology reports will be discussed. Finally, we will consider the recent evolution of this field and the efforts for community-based evaluations.

The search for novel drug targets relies, essentially, on computational methods that prioritize proteins based on inferences from sequence and structure similarities, commonly followed by time-consuming manual examination of information contained in databases and biomedical literature.

Large-scale experimental techniques such as microarrays, two-hybrid systems, protein chips and complex purification methods provide large data collections and add a rich source of additional information. However, the increased volume, complexity and variety of data also generate additional complications for their interpretation.

All of these methods require detailed analysis by experts in the field. Current knowledge of protein function is based on extrapolation of the information accumulated for a relatively small set of proteins for which direct functions have been determined experimentally (<10% of the proteins in well-annotated databases such as SwissProt).

Automated analysis of protein function has additional limitations because protein function is less conserved than protein sequence, and annotations and descriptions in databases do not necessarily reflect all of the available information about protein function [1].

A new generation of applications aims to assist researchers in obtaining and managing additional information by incorporating text-mining and natural-language processing (NLP) tools for the extraction and compilation of functional characteristics of individual genes and proteins.

Furthermore, there is increasing interest in linking unstructured data extracted from free text to information stored in genome and annotation databases such as SwissProt [2] and the *Saccharomyces* Genome Database (SGD) [3]. In this article, we will address some of the methods employed in the processing of complex textual information and discuss their application to the field of bioinformatics and drug discovery (for additional reviews relating to text mining and NLP in the biomedical and molecular biology domain, see Refs [4–6]).

Information resources for text mining

Text-mining applications integrate a broad spectrum of heterogeneous data resources, providing tools for the analysis, extraction and visualization of

Martin Krallinger
Alfonso Valencia*
Protein Design Group, National Center of Biotechnology (CNB-CSIC), Cantoblanco, E-28049 Madrid, Spain
Ramon Alonso-Allende Erhardt
Bioalma SL, Ronda de Poniente 4 – 2nd floor, Unit C-D, Tres Cantos, 28760 Madrid, Spain
*e-mail: valencia@cnb.uam.es

1359-6446/04/$ – see front matter PII: S1359-6446(04)03376-3

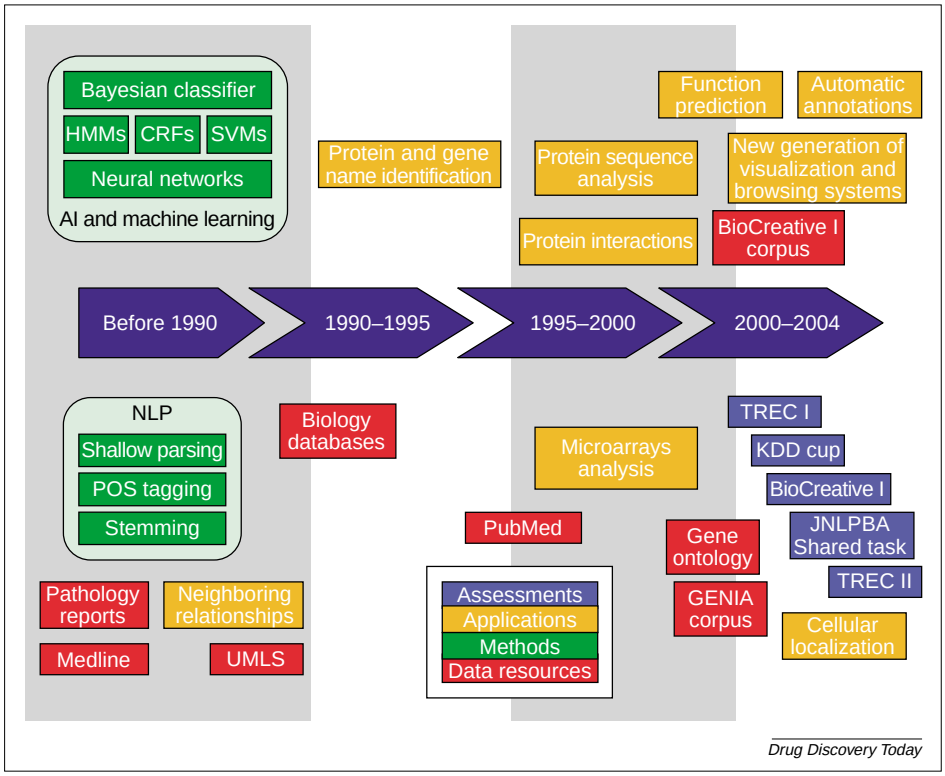

FIGURE 1

Historical perspective of the use of NLP in biomedicine and molecular biology. The hits are divided into different categories: dark-blue boxes show the different community-wide evaluations, whereas orange boxes refer to applications of text-mining strategies in biomedicine and molecular biology. Methods used for text mining and information extraction, such as artificial intelligence (AI), ML and statistical NLP techniques, are shown in green boxes, whereas relevant data resources are depicted in red boxes. Abbreviation: CRF, conditional random fields.

information, with the aim of helping biologists to transform available data into usable information and knowledge.

In molecular biology, the information resources available are, essentially, a vast collection of databases that covers a broad range of source types such as keywords, protein sequences, abstracts and structural information [2]. In addition, some databases focus on specific aspects of protein function, as in the case of protein-interaction databases such as INTACT [7], BIND [8], DIP [9] and MINT [10], whereas others focus on model organisms such as yeast (SGD) [3], *Drosophila* [11] or mouse [12]. The primary source of free textual data information in molecular biology and biomedicine is Medline, which is a collection of more than 12 000 000 abstracts maintained by the National Library of Medicine (NLM) [13] that is commonly accessed by biologists using the PubMed suite.

The main entry points in biological databases are genes and proteins; consequently, the databases contain lists of the names and symbols of genes and proteins, and many of their synonyms. These lists are commonly used as dictionaries for text-mining tools, to index documents and tag genes within free text. The keywords and annotations manually linked to this information in the databases are also used as the benchmark for text-mining and information-extraction tools. In addition, ontologies and thesauri have been developed to classify concepts in protein function, providing a formal framework for the representation of knowledge. A range of text-mining and knowledge-discovery tools has been developed to make use of these classifications.

Gene ontology (GO) [14] is the most widely used classification in molecular biology. The principal use of the concepts encompassed by GO has been in the manual annotation of proteins from model organisms [15]. Several experiments have used GO as a benchmark for text-mining and information-extraction approaches. The drug ontology developed by the Manchester Medical Informatics Group [16] is another ontology applied in biomedicine.

For the analysis of medical text using NLP approaches, medical ontologies such as the Unified Medical Language System (UMLS) metathesaurus [17], with more than 2 000 000 names for ~900 000 concepts (as of June 2004), is of special importance. UMLS was developed as a standard for medical pathological terms [18]. Currently, it integrates ~8 800 000 concepts from 100 biomedical terminologies. Among the resources used for indexing PubMed articles are the Medical Subject Headings (MeSH) terms, which are composed of controlled vocabulary, and the National Cancer Institute ontology of cancer concepts [19].

The development of text-mining technology, largely based on automated learning methods, depends crucially on the availability of repositories of properly annotated text. Currently, the field has a few data corpora available that are widely used. The GENIA corpus [20] is a collection of semantically annotated documents principally related to transcription factors and human blood cells. This corpus has been used in approaches such as the shared task (bioentity recognition) of the JNLPBA workshop [21]. A different dataset was obtained through the BioCreative evaluation [22]. This dataset contains a large collection of text passages related to protein function (GO terms), including their manual classification by the GO curators [23].

Text mining and NLP

The field of NLP is concerned with the analysis of free textual information and has been applied recently in the context of molecular biology. Text-mining approaches involve analyzing and extracting information from large collections of free textual data by using automatic or semi-automatic systems. Currently, text-mining applications are being employed in the identification of biological entities such as protein or gene names, automated protein

annotation, analysis of microarrays and extraction of protein–protein interactions (Figure 1).

In general, text-mining applications take advantage of a range of domain-independent methods such as part-of-speech (POS) taggers, which label each word with its corresponding part of speech (e.g. noun, verb or adjective), or stemmers, which are algorithms that return the morphological root of a word form. Also, domain-specific tools and resources such as protein taggers and ontologies are employed.

Tagging biological entities

The identification of entity types (e.g. company names and places) in textual data is known as 'named entity recognition' or 'semantic tagging', and has been an area of interest in NLP for many years. In biomedical literature, the identification of biological entities such as gene and protein names, chemical compounds and diseases is crucial for facilitating the retrieval of relevant documents and the identification of relationships between those biological entities (e.g. between proteins and diseases). Biomedical language and vocabulary is highly complex and rapidly evolving, making the identification of entities a cumbersome task, especially in the case of protein and gene names. When labeling text relative to the occurrence of genes or proteins, several obstacles are encountered [24]. First, a variety of alternative expressions that refer to the same protein object are often encountered; proteins might be mentioned in documents in terms of gene symbols, protein names, synonymous gene names and typographical variants. Moreover, some gene symbols are ambiguous and might correspond to disease names or experimental methods. The only way to tag these genes is by taking into account the context in which they are referred to. For example, 'EGFR' might correspond to 'epidermal growth factor receptor' or 'estimated glomerular filtration rate', depending on the context. A range of different approaches for handling this problem has been developed, and various community-wide assessments have been carried out to estimate the accuracy of such tools.

Among the strategies adopted to tag proteins and genes are methods such as *ad hoc* rule-based approaches [25], approaches using dictionaries of genes with subsequent exact or inexact pattern matching [26], various machine-learning (ML) techniques and hybrid approaches that take advantage of different techniques [27]. ML techniques refer to statistical and probabilistic models that estimate dependencies between data to make predictions. In this context, support vector machines (SVMs) [28] and hidden Markov models (HMMs) [29] have been applied. The use of naive Bayesian learning, decision trees and inductive rule learning has also been explored [30].

Several approaches to the problem of chemical-name identification have been implemented [31], with one of the main difficulties being the conjunctive nature of the names (i.e. several concepts are contained within a single unbroken string).

Information retrieval of biomedical articles

Information retrieval (IR) is concerned with the recovery of textual information from document collections (e.g. all the documents relevant to a certain protein or disease). In the biomedical domain, IR technologies are in widespread use. Most experimental biologists take advantage of the PubMed information-retrieval system available at the NCBI, which runs on the PubMed database [32]. This system incorporates simple Boolean query searches based on indexed look-up techniques, and a document-similarity search engine based on word-frequency similarities (word-vector neighboring relationships) [33].

Information extraction

Information extraction attempts to identify biologically meaningful semantic structures within free text using strategies based on POS information, ontologies or the identification of patterns. An example of the use of information-extraction applications in molecular biology is the identification of protein interactions.

In the biomedical domain, extracted entities often correspond to proteins, genes, diseases or chemical compounds, for which automated identification methods are often incorporated. For the extraction of entities (in addition to relationships between entities of interest), parsing tools and POS taggers that can detect verbs of interest are also often useful [34].

Knowledge discovery

The volume of scientific literature makes it increasingly difficult to focus on relevant information. Techniques such as pattern matching and syntactic analysis can highlight relevant text passages from large abstract collections. However, generating new insights to direct future research is far more complex. The goal of knowledge discovery is to find hidden information in the literature by exploring the internal structure of the knowledge network created by the textual information. Knowledge discovery could be of major help in the discovery of indirect relationships, which might imply new scientific discoveries. Such new discoveries might provide hints for experts working on specific biological processes.

Applications of text mining

Functional annotation

Annotation of the function of genes and proteins is the principal goal of genome analysis. Classical computational approaches relied on protein-sequence similarity and database annotations. A typical example is the EUCLID system [35] for the classification of proteins into functional groups based on SwissProt database keywords. Other systems [36–39] rely on rules for transferring database information according to the relationships between proteins in families.

Information-extraction techniques have been developed with the aim of obtaining information that is not immediately available from biological databases. For example,

Andrade *et al.* [40] developed one of the first systems in this area by detecting terms in the scientific literature that are statistically associated with literature linked to protein families.

Although keyword-based approaches can cover varying degrees of functional description, they have extremely limited expressivity. Thus, other approaches use ontologies such as GO as a better way of structuring knowledge. For example, Raychaudhuri *et al.* [41] explored the use of different document-classification methods for this task, and Xie *et al.* [42] combined sequence-similarity scores and textual information to support functional annotation using GO.

Along similar lines, text mining has also been used to assist the identification of remote homolog proteins by combining similarity scores and document similarity [43].

Cellular location

Protein activity is associated with specific cellular environments. Several experimental techniques can determine the subcellular localization of a protein, and several recent studies have addressed the extraction of this information from the literature. For instance, Nair and Rost [44] exploited lexical information present in annotation database records to predict the location of proteins, and Stapley *et al.* [45] used a system based on SVMs to classify proteins according to their subcellular localization, extracted from PubMed abstracts.

DNA-expression arrays

Data generated from expression-array experiments are increasing in both volume and complexity. The corresponding analyses focus on the statistical detection of groups of genes with similar expression patterns. Literature-analysis tools provide an alternative insight into the interpretation of array experiments by enabling analysis of the statistical properties of the words present in the abstracts that are associated with genes displaying similar expression patterns (gene clusters). Oliveros *et al.* [46] and Blaschke *et al.* [47] developed the GEISHA method, which uses this type of statistical approach. Shatkay *et al.* [48] also used statistical methods to extract characteristic content-bearing terms for a set of gene-associated documents. Thus, statistical analysis of gene-indexed articles might be useful for the extraction of relevant words and terms for gene clusters.

A new perspective on the problem has been adopted by Raychaudhuri *et al.* [49], who quantified the difference between terms associated with different gene clusters by scoring them according to the functional coherence of the corresponding gene group.

Other approaches use manually annotated keywords or concepts (derived, for instance, from GO) to the expression-array genes to analyze which concepts are relevant to the different gene groups. The FatiGO system [50] extracts relevant GO terms for a group of genes, with respect to the reference set of genes. In the case of the PubGene [51] system, the analysis of microarray data is based on a previously constructed literature network for human genes that are linked to terms from the MeSH database and GO.

Common problems associated with all of these statistical approaches include the unequal distribution of genes in clusters and the number of publications associated with the genes. The non-homogeneous distribution of functional references in the corresponding articles is also an issue.

Protein interactions

High-throughput experiments can generate large-scale protein-interaction networks such as the recently published map of interactions for the *Drosophila* genome [52], thus constituting an amazing new source of information about protein function and potential new drug targets. Information-extraction methods are well positioned to participate in the analysis of this information by connecting the new experiments to the information previously accumulated in the literature, complementing bioinformatics approaches for the prediction of protein interactions [53].

Syntactic predicational structures and semantic propositions referring to binding relationships were used by Rindflesch *et al.* [54] to extract macromolecular-binding terminology. Blaschke *et al.* [55] developed an approach that encapsulated representative relationships between proteins in common descriptions, called 'frames'. Examples of such frames are 'protein X binds to protein Y' and '…complex between protein A and protein B'. The effectiveness of each of the frames was evaluated against a large data collection [56] and embedded in a visualization, analysis and manipulation system for the representation of the network (the SUISEKI system [57]).

Ng *et al.* [58] developed a similar, rule-based model with which to detect protein-activation or -inhibition relationships. Ono *et al.* [59] developed a system to handle long phrases in the literature related to *Escherichia coli*. The method is based on word patterns and manually established POS rules. Recent approaches apply dynamic programming to mine automatically for verbs in sentences in which protein names have been identified previously [60].

Donaldson *et al.* [61] constructed PreBIND and Textomy – an information-extraction system that uses SVMs to evaluate the importance of protein–protein interactions.

More recently, Hoffmann *et al.* [62] implemented a new public server to facilitate access to protein-relationship extractions from the literature (iHOP). Here, the presence of protein names in text sentences is used to hyperlink the corresponding articles, and the densely connected network created by the ubiquitous presence of gene names in scientific abstracts enables fast navigation between different areas of the literature. The incorporation of this concept, together with database and graphical facilities, makes iHOP the first open-access large-scale system for literature navigation based on the concept of protein interactions.

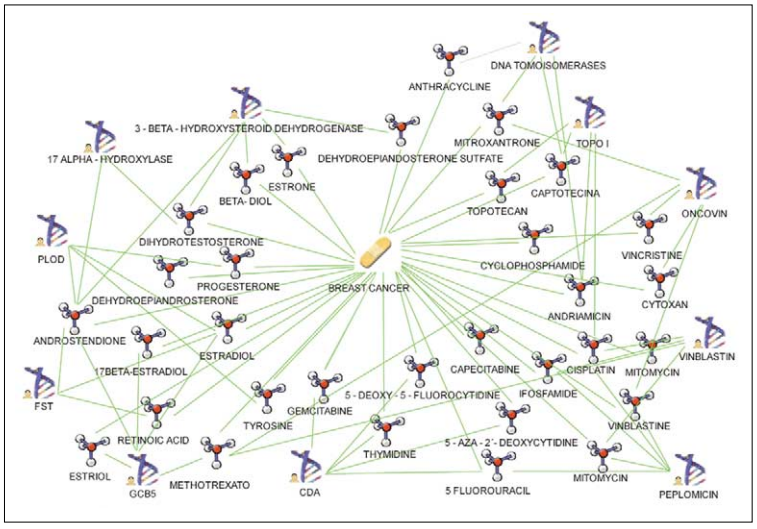

FIGURE 2

Sample output of Alma Knowledge Discovery system. 'Breast cancer' is used as an initial query concept. All the disease-related compounds and their respective relationships with genes are displayed. This visualization enables the discovery of new insights regarding associations between breast-cancer-related genes. They are either associated through a common compound or a new relationship to breast cancer is shown directly. Some of them might have relevant implications for the disease and can be considered for a more detailed study.

Molecular medicine

Several text-mining, NLP and knowledge-discovery applications have been developed for the biomedical domain. They include tools that discover new relationships between relevant entities such as chemical substances and diseases, NLP systems that extract and structure information contained in clinical records, and systems that identify and visualize interactions between molecular substances of interest.

Automated textual analysis to discover new, so-called 'undiscovered', public knowledge and to test proposed hypotheses was first carried out by Swanson and colleagues. They established an indirect connection between dietary fish oil and the circulatory disorder known as Raynaud's disease [63]. A collection of articles on both subjects was available at the time, but no suggestions about using dietary fish oil to treat this disease had been proposed before. The starting point of such approaches is a given concept 'A' (e.g. disease, compound or gene). If A is associated with a second concept 'B', and B is related to a third entity 'C', then A might be related to C even though there is no direct association between them. The proposed relationship should then be further confirmed or rejected using human judgment, laboratory methods or clinical investigations, depending on the nature of the concepts (Figure 2). This principle was also adopted by others to extract indirect associations between estrogen and Alzheimer's disease [64]. Such systems have been used not only to propose new therapeutic strategies but also to extract potentially adverse drug effects or even animal models for certain human disorders.

NLP systems have also been constructed to aid the processing of clinical information contained in medical records. The evaluation of these systems is especially delicate owing to data-privacy issues. MedLEE [65] is a system that processes medical records to extract and structure clinical information, and has been used for years by the New York Presbyterian Hospital Clinical Information System. The GENIES approach [66], contained in the integrated GeneWays system [67], carries out automatic analysis, and extraction of molecular-interaction data and pathways from full-text journals.

Of special interest for drug discovery are relationships between genes and drugs in the context of certain pathologies. This has been addressed by Rindflesch *et al.* [68], who developed EDGAR, a system that extracts relationships between drugs and genes involved in cancer using syntactic analysis and UMLS terms.

Several commercial biomedical-text analysis platforms are currently available. Some of them have been developed directly by pharmaceutical companies, such as the Novartis Knowledge Space Portal (http://www.novartis.com/). Also, bioinformatics companies have constructed biomedical-text-mining applications such as the Alma Knowledge Discovery system (http://www.almabioinfo.com/) (Figure 2), which incorporates powerful database systems, version control, security systems and integrated representation mechanisms. There are also other commercial text-mining and knowledge-discovery applications, including Biovista (http://www.biovista.com/), BioWisdom (http://www.biowisdom.com/), SAS® Text Miner (http://www.sas.com/technologies/analytics/datamining/textminer/) and TextSense (http://www.inforsense.com/products/textsense.html). Biovista, for example, exploits the use of different views or representations of biological knowledge, taking into account context information, and can extract interactions between genes and proteins from free text. BioWisdom uses an extensive ontology of pharmaceutically relevant concepts within its knowledge-discovery platform.

Evaluation of text-mining strategies

As in other areas of text mining and bioinformatics, the field is still in an early phase and, therefore, these and other developments will benefit greatly from the availability of open standards and community-accepted evaluations (Figure 1). For instance, the text-retrieval conference (TREC) genomics track [69] was concerned with the evaluation of *ad hoc* retrieval and information-extraction approaches applied to biological articles. The aim of the knowledge-discovery and data-mining (KDD) challenge cup [70] was to study how text-mining tools can assist biological databases by evaluating how they can support the process of database curation (in this case, the FlyBase database). The identification of biological entities (e.g. proteins and genes) using the GENIA corpus was evaluated at the JNLPBA shared task [21].

The BioCreative evaluation [22] comprised a more biologically inspired evaluation in which biologically

meaningful tasks were prioritized. Both the problem of extraction and normalization of protein and gene names in scientific texts [71] and the extraction of protein annotations from full-text scientific articles were addressed. Both subtasks resulted in the BioCreative corpus, which serves as a gold standard with which to train and test biomedical-text-mining tools. The combination of sentence-classification and pattern-matching techniques, and the use of the information content associated with the words that form query concepts seem promising for the achievement of high precision and recall, respectively, in the second BioCreative task. The combination of different ML techniques obtained good results in the first BioCreative subtask and the JNLPBA shared task.

The future of biomedical-text mining

The increasing interest in the unification of efforts in biomedicine and molecular biology will require access to well-established text sources and data repositories. Other areas in which concerted effort will be required are the development of evaluation systems, the organization of common standards and the organization of the community in the face of common challenges that have been a key factor in the rapid development of text mining in molecular biology and other areas of information extraction. Similar efforts will be required in the domain of molecular medicine to focus community efforts to take advantage of the possibilities provided by the databases and text sources available in molecular biology. In the future, biomedical-text mining might provide new approaches for drug discovery that exploit efficiently indirect relationships derived from bibliographic analysis of entities contained in biological databases (e.g. genes, proteins and chemical compounds).

References

1 Devos, D. and Valencia, A. (2000) Practical limits of function prediction. *Proteins* 41, 98–107
2 Boeckmann, B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370
3 Dwight, S.S. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* 30, 69–72
4 Blaschke, C. *et al.* (2002) Information extraction in molecular biology. *Brief. Bioinform.* 3, 154–165
5 Shatkay, H. *et al.* (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.* 10, 821–855
6 Hirschman, L. *et al.* (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 18, 1553–1561
7 Hermjakob, H. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32 Database issue, D452–D455
8 Bader, G.D. *et al.* (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31, 248–250
9 Xenarios, I. *et al.* (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305
10 Zanzoni, A. *et al.* (2002) MINT: a Molecular INTeraction database. *FEBS Lett.* 513, 135–140
11 FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* 31, 172–175
12 Blake, J.A. *et al.* (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.* 31, 193–195
13 Wheeler, D.L. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31, 28–33
14 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
15 Camon, E. *et al.* (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* 32 Database issue, D262–D266
16 Wroe, C.J. *et al.* (2000) A common drug ontology for decision support and reference. *Conf. Proc. Current perspectives in healthcare computing*, 93–102
17 Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32 Database issue, 267–270
18 Lindberg, C. (1990) The Unified Medical Language System (UMLS) of the National Library of Medicine. *J. Am. Med. Rec. Assoc.* 61, 40–42
19 Golbeck, J. (2003) The National Cancer Institute's thesaurus and ontology. *Journal of Web Semantics* 1, 75–80
20 Kim, J.D. *et al.* (2003) GENIA corpus – semantically annotated corpus for bio-textmining. *Bioinformatics* 19, i180–i182
21 Kim, J.D. *et al.* (2004) Introduction to the Biol.-entity Recognition Task of the JNLPBA workshop. *Proc. of the JNPBA*, 70-76
22 Blaschke, C. *et al.* Evaluation of BioCreative assessment of task 2. *BMC Bioinformatics* (in press)
23 Camon, E.B. *et al.* Evaluation of GO annotation retrieval for BioCreative, Task 2: lessons to be learned and comparison with existing annotation techniques in GOA. *BMC Bioinformatics* (in press)
24 Krallinger, M. *et al.* (2004) Assessing the correlation between contextual patterns and biological entity tagging. *Proc. JNLPBA*, 36–42
25 Fukuda, K. *et al.* (1998) Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.* 1998, 705–716
26 Krauthammer, M. *et al.* (2000) Using BLAST for identifying gene and protein names in journal articles. *Gene* 259, 245–252
27 Tanabe, L. *et al.* (2002) Tagging gene and protein names in biomedical text. *Bioinformatics* 18, 1124–1132
28 Kazama, J. *et al.* (2002) Tuning support vector machines for biomedical named entity recognition. *Proceedings of the Natural Language Processing in the Biomedical Domain,* 1–8
29 Nobata, C. *et al.* (1999) Automatic term identification and classification in biology texts. *Proceedings of the Natural Language Pacific Rim Symposium*, 369–375
30 Hatzivassiloglou, V. *et al.* (2001) Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics* 17 (Suppl. 1), S97–S106
31 Wilbur, W.J. *et al.* (1999) Analysis of biomedical text for chemical names: a comparison of three methods. *Proc AMIA Symp.* 1999, 176–180
32 Schuler, G.D. *et al.* (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.* 266, 141–162
33 Wilbur, W.J. and Coffee, L. (1994) The effectiveness of document neighboring in search enhancement. *Inf. Process. Manage.* 30, 253–266
34 Hobbs, J.R. (2002) Information extraction from biomedical text. *J. Biomed. Inform.* 35, 260–264
35 Tamames, J. *et al.* (1998) EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics* 14, 542–543
36 Fleischmann, W. *et al.* (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics* 15, 228–233
37 Abascal, F. *et al.* (2003) Automatic annotation of protein function based on family identification. *Proteins* 53, 683–692
38 Frishman, D. *et al.* (2001) Functional and structural genomics using PEDANT. *Bioinformatics* 17, 44–57
39 Scharf, M. *et al.* (1994) GeneQuiz: a workbench for sequence analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 348–353
40 Andrade, M.A. *et al.* (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14, 600–607
41 Raychaudhuri, S. *et al.* (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.* 12, 203–214
42 Xie, H. *et al.* (2002) Large-scale protein annotation through gene ontology. *Genome Res.* 12, 785–794
43 MacCallum, R.M. *et al.* (2000) SAWTED: structure assignment with text description-enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics* 16, 125–129
44 Nair, R. and Rost, B. (2002) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics* 18 (Suppl. 1), S78–S86
45 Stapley, B.J. *et al.* (2002) Predicting the sub-cellular location of proteins from text using support vector machines. *Pac. Symp. Biocomput.* 2002, 374–385

46 Oliveros, J.C. *et al.* (2000) Expression profiles and biological function. *Genome Inform. Ser. Workshop Genome Inform.* 11, 106–117
47 Blaschke, C. *et al.* (2001) Mining functional information associated with expression arrays. *Funct. Integr. Genomics* 1, 256–268
48 Shatkay, H. *et al.* (2000) Genes, themes and microarrays. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8, 317–328
49 Raychaudhuri, S. *et al.* (2003) A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics* 19, 396–401
50 Al-Shahrour, F. *et al.* (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20, 578–580
51 Jenssen, T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28
52 Uetz, P. *et al.* (2004) Protein interaction maps on the fly. *Nat. Biotechnol.* 22, 43–44
53 Valencia, A. *et al.* (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* 12, 368–373
54 Rindflesch, T.C. *et al.* (1999) Mining molecular binding terminology from biomedical text. *Proc of the AMIA Annual Symposium,* 127–131
55 Blaschke, C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1999, 60–67
56 Blaschke, C. *et al.* (2002) The frame-based module of the Suiseki information extraction system. *IEEE Intell. Syst.* 17, 14–20
57 Blaschke, C. and Valencia, A. (2001) The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform. Ser. Workshop Genome Inform.* 12, 123–134
58 Ng, S.K. and Wong, M. (1999) Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform. Ser. Workshop Genome Inform.* 10, 104–112
59 Ono, T. *et al.* (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics* 17, 155–161
60 Hatzivassiloglou, V. *et al.* (2002) Learning anchor verbs for biological interaction patterns from published text articles. *Int. J. Med. Inf.* 67, 19–32
61 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
62 Hoffmann, R. *et al.* (2004) A gene network for navigating the literature. *Nat. Genet.* 36, 664
63 Swanson, D.R. (1986) Fish oil, Raynaud's syndrome and undiscovered public knowledge. *Perspect. Biol. Med.* 30, 7–18
64 Smalheiser, N.R. *et al.* (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.* 57, 149–153
65 Friedman, C. (2000) A broad-coverage natural language processing system. *Proc. AMIA Symp.* 270–274
66 Friedman, C. *et al.* (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 17, S74–S82
67 Rzhetsky, A. *et al.* (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathways data. *J. Biomed. Inform.* 37, 43–53
68 Rindflesch, T.C. *et al.* (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* 2000, 517–528
69 Hersh, W. *et al.* (2003) TREC GENOMICS track overview. *Proc. Twelfth Text Retrieval Conference,* 14–24
70 Yeh, A.S. *et al.* (2003) Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics* 19 (Suppl. 1), 331–339
71 Hirschman, L. *et al.* Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC bioinformatics* (in press)