Christian Blaschke
is post-doctoral fellow in the Protein Design Group at the Spanish National Centre for Biotechnology. Dr. Blaschke's main scientific interest is the extraction of biologically relevant information from the scientific literature with statistical and linguistic methods and the application of machine learning techniques to reduce the cost of adapting these systems to new domains.

Lynette Hirschman
is Chief Scientist for the Information Technology Division at the MITRE Corp. in Beford, Massachusetts, USA. Her recent research has focused on the intersection of natural language processing and biomedical informatics.

Alfonso Valencia Herrera
is senior scientist of the Spanish research council (CSIC), Coordinator of the Spanish Network of Bioinformatics, an editorial board member of *Bioinformatics*, and Vice-president of the International Association for Computational Biology (ISCB). In 1994 Dr Valencia created a multidisciplinary group of 20 researchers at the Spanish National Centre for Biotechnology. His main scientific interest is the use of the genomic and proteomic information for the study of molecular evolution and for the development of new biotechnological resources.

***Keywords:** information extraction, molecular biology, ontologies, protein–protein interactions, document retrieval*

Alfonso Valencia,
Protein Design Group,
National Center for Biotechnology,
CNB-CSIC, Cantoblanco,
Madrid E-28049, Spain

Tel: +34 91 585 45 70
Fax +34 91 585 45 06
E-mail: valencia@cnb.uam.es

Information extraction in molecular biology

Christian Blaschke, Lynette Hirschman and Alfonso Valencia
Date received (in revised form): 12th April 2002

Abstract
Information extraction has become a very active field in bioinformatics recently and a number of interesting papers have been published. Most of the efforts have been concentrated on a few specific problems, such as the detection of protein–protein interactions and the analysis of DNA expression arrays, although it is obvious that there are many other interesting areas of potential application (document retrieval, protein functional description, and detection of disease-related genes to name a few). Paradoxically, these exciting developments have not yet crystallised into general agreement on a set of standard evaluation criteria, such as the ones developed in fields such as protein structure prediction, which makes it very difficult to compare performance across these different systems. In this review we introduce the general field of information extraction, we outline the status of the applications in molecular biology, and we then discuss some ideas about possible standards for evaluation that are needed for the future development of the field.

INTRODUCTION – NATURAL LANGUAGE PROCESSING

Despite the widespread use of computers in biological research, the end result of almost all scientific experiments is a publication in the form of text and figures and this is unlikely to change in the foreseeable future. Even if standards are developed for the deposition of some of this valuable information in computer-readable form, the problem of retrieving all past knowledge of molecular biology is staggering. There is thus considerable interest in developing methods that can extract at least part of this information from the literature and convert it from free text to a structured form that is computer readable and can help biologists in their analysis of complex biological problems.

This interest is reflected in the growing number of special workshops and conference sessions on natural language processing and information extraction in biology and biomedicine. For example, the Pacific Symposium on Biocomputing (PSB), starting in 2000, has had a special session on text analysis.[1–3] At the International Conference on Intelligent Systems in Molecular Biology (ISMB) 2001, a satellite workshop was dedicated to text mining in biology.[4] This interest has not been confined to biology meetings. This year, the Association for Computational Linguistics will hold a workshop for natural language processing (NLP) in biology and medicine[5] in the framework of its yearly conference (June 2002), and there will be an exploratory 'track' at the annual Text Retrieval Conference (TREC) on Genomics and Text Retrieval. In this context, we review work going on in the various important sub-areas of natural language processing for biology.

The growing interest in applying natural language techniques to the biomedical literature derives from two forces: an urgent need on the part of biologists to find information in the ever-expanding biological literature; and increased success in applying NLP techniques to Web-based information access needs. The major successes to date for NLP technology have been in areas such as news capture and processing.

There is also a long history of research

NLP applications in medicine have a long history

on applications in medicine. Applications to the medical field focus on two distinct sub-problems: improved access to the medical literature and extraction of information from patient records. Research on access to the medical literature overlaps the work on access to the biomedical literature, although the application focus is somewhat different: more information retrieval for clinical questions for the medical literature *v.* more text data mining applications for the biomedical literature. One operational system oriented towards the medical literature is AcroMed;[6] this system decodes acronyms and abbreviations found in MEDLINE.

IE systems compete since the late 1980s in Message Understanding Conferences

For the handling of medical records, the Medical Language Extraction and Encoding System (MedLEE) is a good example of a deployed system based on natural language processing and information extraction. It is being used at Columbia Presbyterian Hospital.[7,8] Another recently described system, MedSynDiKATe,[9] has taken an ambitious approach, combining knowledge-based methods with linguistic processing to acquire knowledge from reports on medical findings (in German). It learns a weak ontology from a nomenclature (UMLS[10]) to create a knowledge base. This knowledge base, coupled with parsing and information extraction techniques, is then used to extract meaning from the medical reports. In general, medical reports and patient records present a somewhat different set of challenges, owing to the different, often telegraphic, style used in these reports.

Precision and recall are about 93–95% for entity recognition, 70–80% for the identification of binary relationships and about 60% for the extraction of complex relationships

Interestingly, the recent progress in NLP has been driven by use of corpus-based and statistical methods. These same methods (hidden Markov models, various machine learning approaches) have been successfully applied by biologists to the analysis of the genome.[11]

Gerard Salton laid the foundations for information retrieval (IR),[12,13] introducing content analysis in the 1960s.[14] He used term weighting,[15,16] which adjusts the weight of a term according to its importance in a document, a procedure that still forms the basis of most document retrieval systems. Because of its long history, IR is a mature technology; state of the art systems can return search results over gigabyte databases in seconds. IR systems have achieved widespread acceptance by making search over large collections possible. Good systems generally provide high precision for the first 10 or so documents, but high sensitivity (recall) is usually very hard to achieve. There has been increased interest in IR with the growth of the Internet. A series of TRECs, focused on comparative evaluation of retrieval systems under varying conditions, has also spurred progress.[17]

Information extraction (IE) is an outgrowth of work in automated natural language processing, which began in the 1950s with work on transformational grammar by Zellig Harris[18,19] and later Noam Chomsky.[20,21] Information extraction technology made rapid progress starting in the late 1980s, thanks to a series of conferences focused on evaluation of IE: the Message Understanding Conferences (MUCs).[22] These techniques reached good levels of precision and recall (93–95 per cent) for identifying entities (eg persons, organisations, locations) in news texts. Precision and recall around 70–80 per cent have been reported for identification of simple binary relations (eg *PERSON located_at LOCATION*). However, extraction of complex events has remained at around 60 per cent balanced precision and recall. An IE system must be designed to extract the entities and relations appropriate to a specific task. Typical tasks have included extraction of information about terrorist attacks (who attacked whom, where and when), or information about corporate acquisitions and mergers. IE systems have also been applied to medical and biological texts, although there are no standard evaluation suites yet for these domains, so it is

IE rules can be learned from training corpora

MEDLINE contains 11 million documents and is steadily growing

difficult to determine whether these domains are easier or harder than their news domain counterparts – however, see Nobata *et al*.[23] for an interesting comparison of extracting person names compared to gene and protein names.

Early extraction systems were built using hand-crafted rules[24,25] but recent developments show that these rules can be learned automatically.[26,27] Statistical techniques have also proved very effective (hidden Markov models, for example) where there are large corpora of training data available.[28] In addition, there has been a move away from rule-based syntactic analysis towards more approximate 'chunking' and partial parsing techniques.[29]

Question answering is a relatively new research area that has arisen in association with TREC.[17] Systems have been able to achieve impressive performance (around 75 per cent correct answers returned for simple factual questions). Systems generally consist of a module that provides an analysis of the question type (eg a 'who' question is looking for a person; a 'when' question is looking for a time), coupled with an IR stage to locate relevant documents or passages, followed by modules for syntactic and semantic analysis of the passage to locate an answer to the question for presentation to the user (Table 1).

MEDLINE AS A SOURCE OF INFORMATION

Access to full-text articles is difficult; each journal has its own organisation and interface and formatting conventions which require the development of hand-crafted rule sets to download the papers. The recent initiation of two projects in the USA and Europe (PubMedCentral[32] and EBioScience[33]) for a centralised store of journal articles and the creation of the computational resources to access distributed repositories with various structures indicate that this situation may change in the future. But fortunately in biology and medicine abstracts are collected and indexed in MEDLINE hosted at the National Library of Medicine (NLM) in Bethesda, MD (MEDLINE[34]). The system at the NLM is called PubMed and indexes 9,741 different journals in Medicine and Molecular Biology. It currently (mid-2001) contains more than 11 million abstracts and is steadily growing (see Figure 1).

Besides the server at the NLM,

Table 1: Areas of research related to the extraction of information from text

Natural language processing (NLP) or text analysis: refers to any technique that makes use of free text. Normally it includes the use of linguistic tools such as a syntactic analyser or semantic classification. NLP is a multidisciplinary field that includes linguistics, computer science, psychology, cognitive science, logic, philosophy among others. Its goal is to create computational models of language that allow computers to 'decode' and interact via natural (human) language.
Information retrieval (IR): deals with the retrieval of relevant documents from a large document collections (or the from Internet via search engines such as Google[30] or Altavista[31]) in response to a user query. The retrieval can be implemented as Boolean keyword retrieval (as in MEDLINE), or using weighted term co-occurrence to compare the query to documents. Retrieval can be enhanced by providing 'seed' documents in addition to the original query.
Information extraction (IE): IE involves the identification of specific predefined classes of entities or relations in text. These entities or relations can be extracted for further automated processing, such as insertion into a database, visualisation, etc. Extraction is also increasingly used in summarisation and even to generate short summaries of articles.
Natural language understanding (NLU): the goal of NLU is for a computer to 'understand' a piece of text (in the sense of interpreting it as human would and acting accordingly). This requires not only knowledge of the syntax or structure of natural language but also 'world knowledge' and semantic interpretation.
Question answering (Q&A): the ability of a system to return answers (not just documents) in response to user queries. Q&A systems typically scan large document collections (or possibly a single large document, such as an encyclopaedia) to locate the answers; it may then either return extracted passages, or it may synthesise a coherent answer from one or more sources. Q&A draws on question analysis (to determine what kind of information is being sought), information retrieval (to locate answer passages), extraction (to identify specific relations) and in some cases, text generation or summarisation, to synthesise answers.

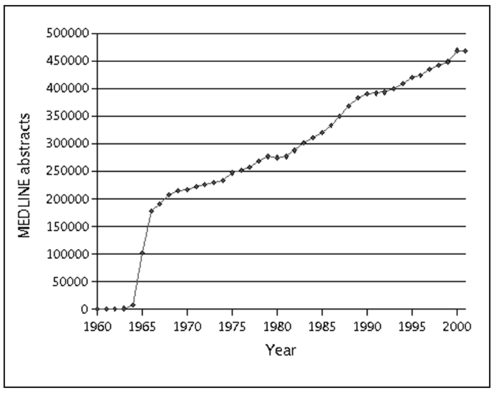

Figure 1: The growth of MEDLINE per year. The figure shows the number of publications that are indexed in MEDLINE from 1960 to the year 2001

MEDLINE abstracts can be collected from the European Bioinformatics Institute (EBI) in Hinxton, Cambridge (SRS[35]), other publicly available MEDLINE servers (DrFelix[36]), or from commercial distributions (SilverPlatter[37]).

APPLICATIONS IN MOLECULAR BIOLOGY

Focus on the technology applied

Statistics of term occurrence

The basic elements of text are words, and their frequencies, co-occurrences and lexical features can be used to cluster and classify text, find documents that treat a similar theme or select significant words that describe a group of documents. One of the earliest applications of these methods in biology was a general text-clustering algorithm developed by Wilbur and Coffee[38] based on word-frequency vectors to find related MEDLINE documents. More specific methods were developed by Andrade and Valencia,[39] who used the characteristics of word distributions in text clusters to extract significant words. The clustering of text based on word distributions was proposed for text classification and organisation of documents.[40,41]

These approaches are limited because words are often ambiguous and refer to more than one object (eg two proteins with the same name). Moreover, different words can have the same meaning (synonyms) and the same word can be part of constructions with very different meanings (eg cell cycle, cell membrane, cell division).

Approaches with deeper syntactical analysis

Methods based on natural language processing (part-of-speech tagging, grammar analysis, analysis of coordination and pragmatics, and natural language understanding) developed in the field of computer science have mostly been applied to the detection of protein–protein and protein–drug interactions (for a discussion and references, see below).

These methods are still limited to relatively small corpora; it is not clear that they will scale up to the millions of abstracts available on MEDLINE, much less to the analysis of the corresponding full text articles. In addition, the use of complex nomenclatures (eg chemical compounds or gene names) will require special sub-grammars.

Mixed approaches

The combination of both term co-occurrence and syntactic approaches has led to significant advances and seems to be highly appropriate for applications in molecular biology. Linguistic tools are good at detecting terms such as the names of proteins, drugs or diseases (with the limitations discussed below in detection of protein and gene names). Statistics on the other hand has been used to describe the relationship between these terms in a

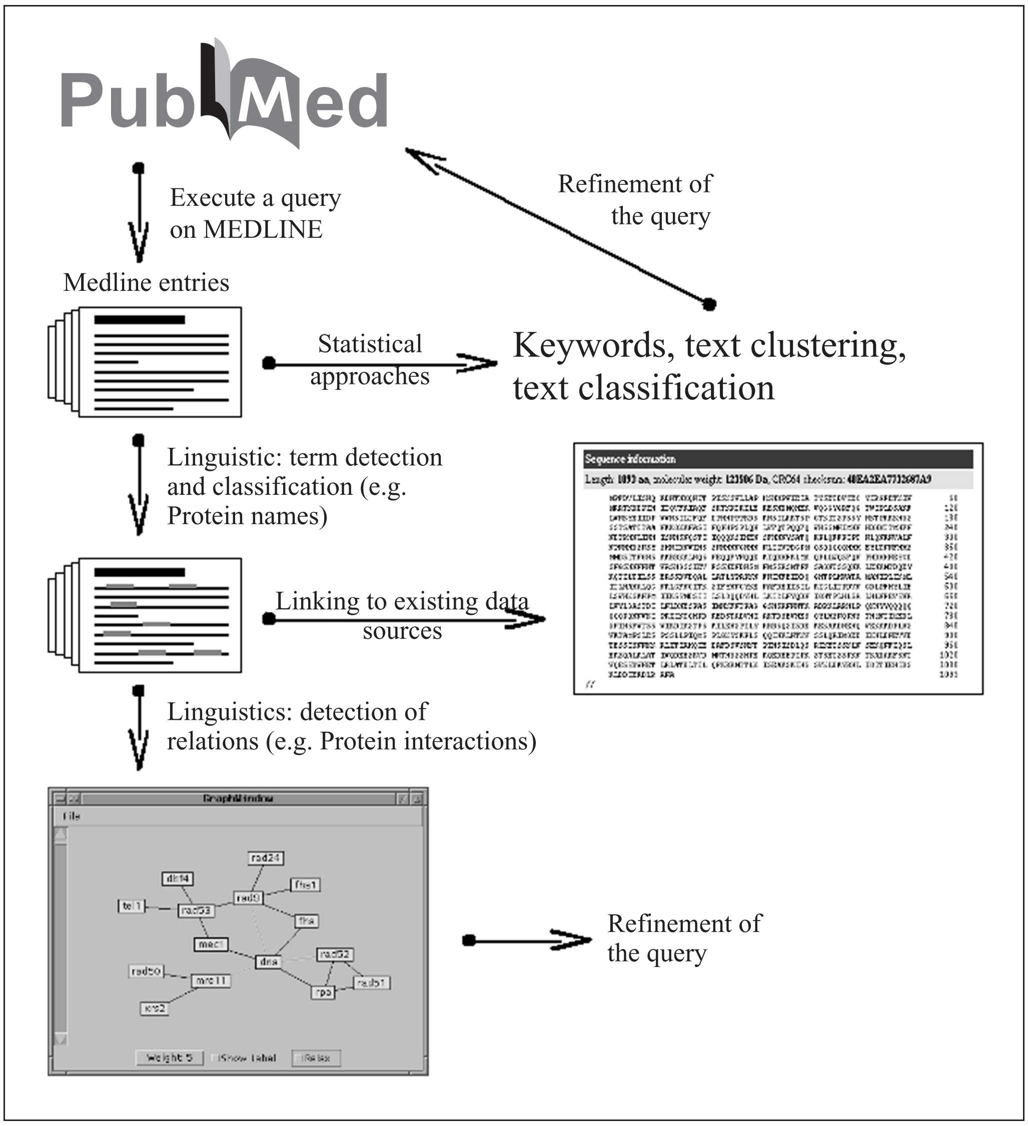

Figure 2: Overview over the general process of information extraction. First the input text is extracted from a document repository (MEDLINE abstracts from PubMed in many cases). Then statistical methods can be applied to extract keywords or classify the documents in predefined classes. Using linguistic approaches, the text can be analysed in more detail, terms such as protein names can be extracted and relations can be detected

probabilistic way what provides great flexibility to this type of systems.[42–45]

Focus on the applications in biology

Extraction of related documents

In biology simpler statistical methods based on word frequencies in documents and more sophisticated linguistic techniques are used

The general goal of IR is to return documents relevant to a user's query on a particular subject or topic of interest. The query can be specified by specific search terms or by an initial set of documents that serves as a sample of relevant documents. In some recent works[38,46] a similarity value, based on the word frequency in abstracts, was used to group 'neighbouring documents' from PubMed together. This helped to expand the set of query terms, to find publications related to the previously selected ones. A limitation of this approach was that it often led to documents that were similar in their word frequencies but not in the content. Recent developments[47] try to overcome this problem and provide a text clustering based on the themes of the documents (that is conceptually a subject area that is discussed by various documents).

The clustering of neighbouring documents is based on the fact that words depend on each other and that documents that have many words in common most likely treat a similar theme. A somewhat more sophisticated treatment of the co-

Applications of NLP systems in biology include:

- **detection of related documents**
- **assignment of protein functions**
- **named entity recognition (especially protein names)**
- **IE for DNA arrays**
- **characterisation of protein localisation**
- **protein–protein and protein–drug interactions**

occurrence levels of words is used by XplorMed[48] to refine Medline queries and reduce the number of unrelated documents in the search results.

A simple keyword extraction system that uses the distribution of words as an indication of their importance was used to find relevant articles for entries in OMIM (Online Mendelian Inheritance in Man[49]) and to keep the literature links up to date.[50]

MedMiner[51] filters and organises textual information and supports the user in retrieving and selecting documents related to a group of genes.

Assignment of protein functions

Proteins are central objects in living systems and the description of their function is one of the key tasks of molecular biology. Therefore it is necessary to extract functional information from the literature to complement the knowledge stored in sequence databases (eg SWISS-PROT[52]).

A method based on the composition of words in protein families (a number of proteins associated by sequence similarity) was AbXstract[39] (Blaschke *et al.*, unpublished), which was born from the need to extract facts related to protein families as part of a system for automatic functional protein sequence analysis. The goal of this system is, for a given sequence family, not to depend entirely on the database annotations but to be able to recover what is published for all the sequences in this family in the literature in the form of keywords and significant sentences selected automatically from the text.

Related to this section are the works of Chang *et al.*[53] and MacCallum *et al.*,[54] who use document similarity scores that indicate the functional relation of proteins to improve the distinction of true and false remote homologues in different types of sequence searches.

Detecting protein names in the literature, and their relation to the database entries

Fukuda *et al.*[55] and Proux *et al.*[56] described the first approaches for extracting protein names from the corresponding noun phrases by part-of-speech taggers and parsers. These noun phrases were analysed with dictionaries and morphological rules.

Leek[57] and Hatzivassiloglou *et al.*[58] used machine learning methods to detect the names and disambiguate them according to their context. Yoshida *et al.*[59] went a step further to find abbreviations (or synonyms) to the names that were detected in the text. This is an extension of the work by Fukuda *et al.*;[55] also see recent work by Pustejovsky *et al.*[60] on decoding acronyms and abbreviations.

The problem of detecting protein and genes names in the literature is intimately related to their mapping to the corresponding database entries. The practical use of this technology in molecular biology cannot be separated from the analysis of experimental results based on genes and proteins, which can be complemented with information extracted from the literature only if the correspondence between literature and database names can be established unambiguously. Blaschke and Valencia[61] have demonstrated that even for human-curated public databases, the correct citation in the literature for the individual items indexed in the database (ie protein interactions) was found only for a small fraction of the entries, mainly because it was impossible to detect the corresponding protein names in the text.

Analysis of expression array experiments

Expression arrays have introduced a paradigmatic change in biology by shifting experimental approaches from single gene studies to genome-level analysis.

Issues related to the first steps of the analysis, including treatment of the DNA chip images and information organisation, have received much attention, including the development of several methods for the identification of groups of genes with similar expression patterns (gene expression clusters[62]). The development of methods to extract information about the common biological characteristics of gene clusters has received considerably less attention. There is an obvious need for protocols to summarise vast amounts of data in a comprehensive way, algorithms to select information that could be of use to human experts, and tools to guide them through the analysis. A similar method to the one used for analysis of protein families[39] was developed to assist in the analysis of DNA expression array experiments (the GEISHA system[63,64]). GEISHA extracts significant parts of the text related to the gene expression clusters by comparing the term frequencies in all the clusters, to aid in the functional analysis of similarly expressed genes.

With a similar goal, Shatkay *et al.*[65] applied a probabilistic method to find general themes within the literature and to extract keywords for each cluster of genes.

Protein localisation

Another important attribute of proteins is their localisation in a cell or a tissue. Craven and Kumlien[66] applied machine learning techniques to extract facts about the sub-cellular or tissue localisation of proteins and their relations to diseases and drugs from which a knowledge base can be constructed. Lexical analysis resulted in the rule-based system Meta_A[67] to classify the entries in the protein database SWISS-PROT in classes of subcellular localisation. Stapley *et al.*[68] demonstrated the efficiency of Support Vector Machines for the prediction of the subcellular localisation of proteins based on term frequencies in their associated MEDLINE abstracts.

Drug–protein interactions

Proteins can interact with chemical substances (metabolite–enzyme interactions) or drugs. EDGAR[69] is, to our knowledge, the only public system that addressed the problem of protein–drug interactions. This system is conceptually very similar to the ones described below, oriented to relationships between proteins. It uses the UMLS Metathesaurus[10] as the primary knowledge source to detect the names of proteins and drugs in the text.

Protein interactions

The problem that has attracted most attention in this field is the retrieval of protein interactions. The solutions range from the simple co-occurrence of gene symbols to methods with a deeper syntactical analysis. A precondition for the detection of protein interactions is the detection of the protein names in the text (see discussion below).

Marcott *et al.*[70] were just interested in retrieving a high number of documents that probably contained information about protein-protein interactions. Stapley and Benoit[71] used fixed lists of gene names and detected relations between these genes by means of co-occurrence in MEDLINE abstracts. Jenssen *et al.*[72] used a similar approach to find relations between human genes and they compared the results to gene clusters obtained from DNA array experiments.

Authors who have followed approaches with a focus on linguistics include: Park *et al.*,[73] who investigated the possible use of Combinatory Categorial Grammar for detecting general relations in biomedical text, and Rindflesch *et al.*,[74] who used biomedical dictionaries (the UMLS MetaThesaurus from the National Library of Medicine) to detect cells and genes in the text and possible relations between them. Sekimizu *et al.*[75] concentrated on frequently seen verbs and the application of a grammar to identify the corresponding subjects and objects of these verbs to detect possible interactions; Thomas *et al.*[76] and Humphreys *et al.*[77] demonstrated the feasibility of adapting a general-purpose information extraction system to the domain of molecular biology, and Yakushiji *et al.*[45] adapted a general-purpose parser and grammar to biomedical text. Similar techniques were applied by Ono *et al.*[78] and Proux *et al.*[79] Friedman *et al.*[80] used a similar NLP technique that was adapted from an earlier

Ontologies are a way to represent knowledge

Widely used ontologies in biology and biomedicine are the UMLs and GO

medical natural language processing system. Pustejovsky *et al.*[81] included the treatment of anaphora which allows the capture of relations across sentence boundaries, an important feature if there is a need to extract all of the relations discussed in a given article.

A pattern-matching method based on constructions that are often found in text combined with limited syntactical analysis is an effective way of extracting information about the type of connection between genes/proteins. This approach is flexible and can be applied to large text corpora. In the first implementation of one such system, Blaschke *et al.*[42] analysed a collection of 100,000 MEDLINE abstracts, avoiding the problem of name detection by assuming a fixed list of protein names. The original application was later improved[42] by the addition of complementary patterns supplementing the first simple heuristic pattern protein_A interaction_verb protein_B and by a module for the automatic detection of protein names based on the analysis of lexical, morphological, syntactical and contextual information. In empirical tests, about 25,000 interactions can be retrieved from 80,000 abstracts related to yeast. The accuracy and recall of their system have been reported to be useful for the biological analysis of the extracted data when a large enough collection of abstracts is used as source of information.

A similar system was described by Ng and Wong[44] and Wong[82] based on the detection of protein names with semantic rules and dictionaries, embedded in an information-retrieval and data-integration system. Unfortunately only the results of the analysis of 26 abstracts have been published. Ono *et al.*[78] present a similar method but they include a limited syntactical analysis to address the problem of coordination.

Regrettably the performances of the different systems cannot be compared since they have been applied to very different text corpora of different sizes with different assumptions about the extraction of protein names and different ways of scoring errors.

Knowledge representation – ontologies

Knowledge representation is integrally related to information extraction. Indeed, information is just the intermediate step between data (the primary result of an experiment) and knowledge (interpretation and conclusions). Information extraction from the literature will be useful only if this information can be related to the existing knowledge. Ontologies are the most common form for the representation of knowledge in the bioinformatics community. An ontology is the specification of the key concepts in a given field and the relations that exist among these concepts. In the simplest case, an ontology is a controlled vocabulary; in more complex scenarios, the relations between the concepts are formulated as axioms that capture the network structure of the knowledge that they model. These axioms can be used to extract implicit knowledge, such as the transitive closure of relations (if an enzyme is a kind-of protein and a protein is a kind-of polypeptide, then an enzyme is a kind-of polypeptide).

Many different ontologies have been developed in the past years.[10,83,84] Two of these have been particularly influential in biology and biomedicine. The first is the Unified Medical Language System (UMLS),[10] which is the largest public repository for terminology in biomedicine. It captures much of the current knowledge and terminology for, eg, diseases, drugs and therapies. It is used for term recognition and classification in many IE applications. The second ontology is from the Gene Ontology Consortium (GO).[84] GO provides a dynamic controlled vocabulary for all organisms that can account for differences between organisms, with sufficient flexibility to accommodate the constant changes in biological knowledge. This initiative has produced considerable interest among the community. It is now being used as an appropriate ‘target structure' for information mining techniques. For example Raychaudhuri *et al.*[85] used machine learning techniques to automatically assign genes mentioned in MEDLINE abstracts to GO concepts.

PERSPECTIVES

In the six years from the first publication on retrieving information from the biology literature, a tremendous interest has grown around these applications. Reviewing the main issues in the field, it is perhaps possible now to separate these into technical and organisational issues. Key technical issues are the identification of protein and gene names, and, very importantly, their relation to the corresponding sequence database entries. Another technical issue concerns the proper combination of linguistic and statistical methods. The main organisational issue is the lack of a common evaluation for the different systems and technologies, with the associated detrimental consequences for the field, both at the scientific and commercial levels. A community 'Challenge Evaluation', similar to the ones developed in the natural language processing and protein structure prediction communities, will require agreement on a problem of practical importance to the biology community, eg extraction of protein interactions, and a well-defined evaluation standard. This will allow researchers to measure the ability of a variety of systems for retrieving information, using all available text resources.

The lack of accepted standards for evaluating IE systems in biology hinders progress in the field

References

1. Hirschman, L., Park, J. C., Tsujii, J. *et al.* (2002), 'Session introduction', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 323–325.
2. Tsujii, J. and Wong, L. (2001), 'Session introduction' in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 372–373.
3. Tsunoda, T. and Wong, L. (2000), 'Session introduction', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 488–489.
4. Text in Biology (2001), 'Biological Research with Information Extraction & Open-Access Publications (BRIE & OAP)', Copenhagen, Denmark, July (URL: http://bioinformatics.org/bof/brie-oap-01/).
5. Natural Language Processing in the Biomedical Domain at the ACL-02 anniversary meeting. University of Pennsylvania, Philadelphia, July 2002 (URL: http://www.acl02.org/).
6. Pustejovsky, J., Castaño, J., Cochran, B. *et al.* (2001), 'Extraction and disambiguation of acronym–meaning pairs in Medline', in 'Proc. of Medinfo', London, September 2001; see also URL: http://www.medstract.org/to download AcroMed.
7. Friedman, C., Alderson, P., Austin, J. *et al.* (1994), 'A general natural language text processor for clinical radiology', *J. Amer. Med. Informatics Assoc.*, Vol. 1, pp. 161–174.
8. Friedman, C. (2000), 'A broad coverage natural language processing system', in Overhage, M., Ed., 'Proceedings of the AMIA Symposium', Hanley & Belfus, Philadelphia, pp. 270–274.
9. Hahn, U., Romacker, M. and Schulz, S. (2002), 'Creating knowledge repositories from biomedical reports: The medSynDiKATe text mining system', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 338–349.
10. Humphreys, B. L., Lindberg, D. A., Schoolman, H. M. and Barnett, G. O. (1998), 'The unified medical language system: An information research collaboration', *J. Amer. Med. Inform. Assoc.*, Vol. 5, pp. 1–11.
11. Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G. J. (1998), 'Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids', Cambridge University Press, Cambridge.
12. Salton, G. (1989), 'Automatic Text Processing: The transformation, analysis and retrieval of information by computer', Addison-Wesley, Reading, MA.
13. Salton, G. (1991), 'Developments in automatic text retrieval', *Science*, Vol. 253, p. 974.
14. Salton, G. (1968), 'Automatic Content Analysis in Information Retrieval', University of Pennsylvania, Philadelphia, PA.
15. Salton, G. and Buckley, C. (1988), 'Term-weighting approaches in automatic information retrieval', *Inf. Proc. Man.*, Vol. 24, pp. 513–523.
16. Salton, G., Wong, A. and Yang, S. S. (1975), 'A vector space model for automatic indexing', *J. Assoc. Comput. Machinery*, Vol. 18, pp. 613–620.
17. Voorhees, E. M. and Harman, D. K. (2000), 'The Ninth Text Retrieval Conference

(TREC-9)', The National Institute of Standards and Technology (NIST), Special Publication 500–249 (also available at URL: http://trec.nist.gov/pubs.html).

18. Harris, Z. (1952), 'Discourse analysis', *Language*, Vol. 28, pp. 18–23.
19. Harris, Z. (1957), 'Co-occurrence and transformation in linguistic structure', *Language*, Vol. 33, pp. 283–340.
20. Chomsky, N. (1956), 'Syntactic Structures', Mouton and Co., The Hague and Paris.
21. Chomsky, N. (1965), 'Aspects of the Theory of Syntax', MIT Press, Cambridge, MA.
22. Hirschman, L. (1998), 'The evolution of evaluation: Lessons from the message understanding conferences', *Computer Speech Language*, Vol. 12, pp. 281–305.
23. Nobata, C., Collier, N. H. and Tsujii, J. (2000), 'Comparison between tagged corpora for the named entity task', in Kilgarriff, A. and Berber Sardinha, T., Eds, 'Proceedings of the Workshop on Comparing Corpora' (at ACL '2000), Hong Kong University of Science and Technology, 7th October, pp. 20–27.
24. Lehnert, W., McCarthy, J., Soderland, S. *et al.* (1993), 'UMASS/HUGHES: description of the CIRCUS system used for MUC-5' in 'Proceedings 5th Message Understanding Conference', Baltimore, MD, pp. 277–291.
25. Riloff, E. and Lehnert, W. (1994), 'Information extraction as a basis for high-precision text classification', *ACM Trans. Inf. Syst.*, Vol. 12, pp. 296–333.
26. Riloff, E. (1996), 'Automatically generating extraction patterns from untagged text', in 'Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)', AAAI Press, Menlo Park, CA, pp. 1044–1049.
27. Soderland, S., Fisher, D. and Lehnert, W. (1997), 'Automatically Learned vs. Hand-crafted Text Analysis Rules', CIIR Technical Report, TE-44.
28. Bikel, D. M., Miller, S. Z., Schwartz, R. and Weischedel, R. (1997), 'Nymble: A high-performance learning name-finder', in 'Proceedings of the Fifth Conference on Applied Natural Language Processing', Washington, DC, 31st March to 3rd April, Association for Computational Linguistics, pp. 194–201.
29. Collins, M. (1997), 'Three generative, lexicalised models for statistical parsing', in 'Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL) and 8th Conference of the European chapter of the Association for Computational Linguistics', Madrid, Morgan Kaufmann/ACL.
30. URL: http://www.google.com
31. URL: http://www.altavista.com
32. A digital archive of life sciences literature managed by the National Center for Biotechnology Information (NCBI). URL: http://www.pubmedcentral.nih.gov
33. The electronic publication initiative at EMBO. URL: http://www.e-biosci.org/
34. URL: http://www.ncbi.nlm.nih.gov/pubmed or http://www.nlm.nih.gov/Entrez/medine.html
35. URL: http://srs.ebi.ac.uk/
36. URL: http://www.beaker.iupui.edu/drfelix/
37. URL: http://www.silverplatter.com
38. Wilbur, W. J. and Coffee, L. (1994), 'The effectiveness of document neighboring in search enhancement', *Inf. Process Man.*, Vol. 30, pp. 253–266.
39. Andrade, M. A. and Valencia, A. (1998), 'Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families', *Bioinformatics*, Vol. 14, pp. 600–607.
40. Iliopoulos, I., Enright, A. J. and Ouzounis, C. (2001), 'TEXTQUEST: Document clustering of MEDLINE abstracts for concept discovery in molecular biology', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 374–383.
41. Renner, A. and Aszodi, A. (2000), 'High-throughput functional annotation of novel gene products using document clustering', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 54–65.
42. Blaschke, C., Andrade, M. A., Ouzounis, C. and Valencia, A. (1999), 'Automatic extraction of biological information from scientific text: Protein–protein interactions', in 'Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 60–67.
43. Blaschke, C. and Valencia, A. (2002), 'The frame-based module of the Suiseki information extraction system', *IEEE Intelligent Systems*, March/April, pp. 14–20.
44. Ng, S. and Wong, M. (1999), 'Toward routine automatic pathway discovery from on-line scientific text abstracts', *Genome Informatics Ser.*, pp. 104–112.
45. Yakushiji, A., Tateisi, Y., Miyao, Y. and Tsujii, J. (2001), 'Event extraction from biomedical papers using a full parser', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 408–419.
46. Wilbur, W. J. and Yang, Y. (1996), 'An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts', *Comput. Biol. Med.*, Vol. 26, p. 209.
47. Wilbur, W.J. (2002), 'A thematic analysis of the AIDS literature', in 'Pacific Symposium on

Biocomputing', World Scientific, River Edge, NJ, pp. 386–397.

48. Perez-Iratxeta, C., Bork, P. and Andrade, M. A. (2001), 'XplorMed: A tool for exploring MEDLINE abstracts', *Trends Biochem. Sci.*, Vol. 26, pp. 573–575.

49. McKusick, V. (1994), 'Mendelian Inheritance in Man, Catalog of Human Genes and Genetic Disorders', John Hopkins University Press, Baltimore, MD.

50. Andrade, M. A. and Bork, P. (2001), 'Automatic extraction of information in molecular biology', *FEBS Lett.*, Vol. 476, pp. 12–17.

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

52. Bairoch, A. and Apweiler, R. (2000), 'The SWISS-PROT protein sequence data bank and its supplement TREMBL', *Nucleic Acids Res.*, Vol. 28, pp. 46–48.

53. Chang, J. T., Raychaudhuri, S. and Altman, R. B. (2001), 'Including biological literature improves homology search', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 374–383.

54. MacCallum, R. M., Kelley, L. A. and Sternberg, M. J. (2000), 'SAWTED: Structure assignment with text description – enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons', *Bioinformatics*, Vol. 6, pp. 125–129.

55. Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T. (1998), 'Information extraction: Identifying protein names from biological papers' in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 707–718.

56. Proux, D., Rechenmann, F., Julliard, L. *et al.* (1998), 'Detecting gene symbols and names in biological texts: A first step toward pertinent information extraction', in 'Genome Informatics Ser.,' Universal Academy Press, Tokyo, pp. 72–80.

57. Leek, T.R. (1997), 'Information extraction using hidden Markov models', Master thesis, University of California, San Diego.

58. Hatzivassiloglou, V., Duboue, P. A. and Rzhetsky, A. (2001), 'Disambiguating proteins, genes, and RNA in text: a machine learning approach', in 'Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 97–106.

59. Yoshida, M., Fukuda, K. and Takagi, T. (2000), 'PNAD-CSS: A workbench for constructing a protein name abbreviation dictionary', *Bioinformatics*, Vol. 16, pp. 169–175.

60. Pustejovsky, J., Castano, J., Cochran, B. *et al.* (2001), 'Automatic extraction of acronym–meaning pairs from MEDLINE databases', in 'Proc. of Medinfo', London, September 2001.

61. Blaschke, C. and Valencia, A. (2001), 'Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study', *Comp. Funct. Genomics*, Vol. 2, pp. 196–206.

62. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 14863–14868.

63. Blaschke, C., Oliveros, J. C. and Valencia, A. (2000), 'Mining functional information associated to expression arrays', *Funct. Integr. Genomics*, Vol. 4, pp. 256–268.

64. Oliveros, J. C., Blaschke, C., Herrero, J. *et al.* (2000), 'Expression profiles and biological function', *Genome Informatics Ser.*, pp. 106–117.

65. Shatkay, H., Edwards, S., Wilbur, W. J. and Boguski, M. (2000), 'Genes, themes, and microarrays. Using information retrieval for large-scale gene analysis', in 'Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 317–328.

66. Craven, M. and Kumlien, J. (1999), 'Constructing biological knowledge bases by extracting information from text sources', in 'Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology' AAAI Press, Menlo Park, CA, pp. 77–86.

67. Eisenhaber, F. and Bork, P. (1999), 'Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries', *Bioinformatics*, Vol. 15, pp. 528–535.

68. Stapley, B. J., Kelley, L. A. and Sternberg, M. J. E. (2002), 'Predicting the sub-cellular location of proteins from text using support vector machines', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 374–385.

69. Rindflesch, T. C., Tanabe, L., Weinstein, J. N. and Hunter, L. (2000), 'EDGAR: Extraction of drugs, genes and relations from the biomedical literature', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 515–524.

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

71. Stapley, B. J. and Benoit, G. (2000), 'Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts', in 'Pacific

Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 529–540.

72. Jenssen, T. K., Laegreid, A., Komorowski, J. and Hovig, E. (2001), 'A literature network of human genes for high-throughput analysis of gene expression', *Nature Genetics*, Vol. 28, pp. 21–28.

73. Park, J. C., Kim, H. S. and Kim, J.J. (2001), 'Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 396–407.

74. Rindflesch, T. C., Hunter, L. and Aronson, A. R. (1999), 'Mining molecular binding terminology from biomedical text', in 'Proceedings of the AMIA Symposium', American Medical Informatics Association, Bethesda, MD, pp. 127–131.

75. Sekimizu, T., Park, H. S. and Tsujii, J. (1998), 'Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts', *Genome Informatics Ser.*, pp. 62–71.

76. Thomas, J., Milward, D., Ouzounis, C. *et al.* (2000), 'Automatic extraction of protein interactions from scientific abstracts', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 384–395.

77. Humphreys, K., Demetriou, G. and Geizauskas, R. (2000), 'Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structure', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 502–513.

78. Ono, T., Hishigaki, H., Tanigami, A. and Takagi, T. (2001), 'Automated extraction of information on protein–protein interactions from the biological literature', *Bioinformatics*, Vol. 17, pp. 155–161.

79. Proux, D., Rechenmann, F. and Julliard, L. (2000), 'A pragmatic information extraction strategy for gathering data on genetic interactions', in 'Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology' AAAI Press, Menlo Park, CA, pp. 279–285.

80. Friedman, C., Kra, P., Yu, H., Krauthammer, M. and Rzhetsky, A. (2001), 'GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles', *Bioinformatics Suppl.*, Vol. 1, pp. 74–82.

81. Pustejovsky, J., Castaño, J., Zhang, J. *et al.* (2002) 'Robust relational parsing over biomedical literature: Extracting inhibit relations', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 362–373.

82. Wong, L. (2001), 'A protein interaction extraction system', in 'Pacific Symposium on Biocomputing', World Scientific, River Edge, NJ, pp. 520–531.

83. Stevens, R., Baker, P., Bechhofer, S. *et al.* (2000), 'TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources', *Bioinformatics*, Vol. 16, pp. 184–186.

84. The Gene Ontology Consortium (2000), 'Gene ontology: Tool for the unification of biology', *Nature Genet.*, Vol. 25, pp. 25–29.

85. Raychaudhuri, S., Chang, J. T., Sutphin, P. D. and Altman, R. B. (2002), 'Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature', *Genome Res.*, Vol. 12, pp. 203–214.