Text-based knowledge discovery: search and mining of life-sciences documents

Robert Mack and Michael Hehenberger

Text literature is playing an increasingly important role in biomedical discovery. The challenge is to manage the increasing volume, complexity and specialization of knowledge expressed in this literature. Although information retrieval or text searching is useful, it is not sufficient to find specific facts and relations. Information extraction methods are evolving to extract automatically specific, fine-grained terms corresponding to the names of entities referred to in the text, and the relationships that connect these terms. Information extraction is, in turn, a means to an end, and knowledge discovery methods are evolving for the discovery of still more-complex structures and connections among facts. These methods provide an interpretive context for understanding the meaning of biological data.

***Robert Mack**
IBM, Thomas J. Watson
Research Centre
19 Skyline Drive
Hawthorne
NY 10532, USA
tel: +1 914 784 7830
fax: +1 914 784 6078
*e-mail: robertmack@
us.ibm.com
Michael Hehenberger
IBM Global Life Sciences
Route 100
Building 3
MD3128
Somers
NY 10589, USA

▼ The sequencing of the human genome has greatly increased the pace of life-science research. The knowledge gained creates new challenges for researchers to understand and apply new life-science data. Structured chemical and biological data generated by experimental techniques such as high-throughput sequencing and high throughput screening (HTS), play a major role in the generation of new knowledge. However, biotechnical literature also plays a crucial role in integrating, annotating and communicating experimental results and their implications. This literature is constantly expanding, and researchers struggle to keep up with both the volume of information and the various domains of expertise represented in the literature.

Text-based knowledge discovery tools and methods can help researchers manage this wealth of information, and discover facts, relationships and implications in biomedical literature that can be used to help solve biotechnical problems. Text searching or traditional information retrieval (IR) plays an important role in this discovery process, but one that is increasingly overshadowed by a new generation of information extraction (IE) capabilities. These capabilities are helping researchers discover much more precise, and fine-grained facts and relationships that address specific questions and topics expressed in text information sources. Moreover, text-mining capabilities have an increasing role to play in the broader methods of biomedical knowledge discovery, in combination with data mining, and modeling of biomedical structures and processes.

In this context, therefore, discovery refers to methods for generating and analyzing compilations of text information that serve as a context for interpreting biological data resulting from a wide range of data-generation methods and experiments. These interpretative contexts provide clues for identifying the role of genes and proteins in cell function and in mechanisms of disease dysfunction, and can contribute to identifying potential drug targets for treating disease. As Ng and Wong wrote, ‘The race to a new gene or drug is now increasingly dependent on how quickly a scientist can keep track of the voluminous information online to capture the relevant picture (such as protein–protein interaction pathways) hidden within the latest research articles’ [1]. Biomolecular data are fundamental to knowledge discovery – in particular, drug discovery – but raw sequence and structure data require a context and explanation to be understood, and integrating text and data is fundamental to creating this context [2].

1359-6446/02/$ – see front matter ©2002 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(02)02286-9

Information retrieval

Text searching involves submitting search specifications, based on keywords that searchers believe are contained in documents, to a search engine, which retrieves documents as search results. In the biotechnical domain, MEDLINE is the standard corpus for describing literature, and search tools such as PUBMED (http://www.pubmed.gov) provide text-search access to these abstracts. Text search-engines analyze documents into 'bags of words', and index the documents associated with each word. Different search methods help users specify relationships between combinations of search keywords, and different search algorithms exist for relating search terms to indexed terms. For example, Boolean searching enables searchers to look for documents containing combinations of keywords specified by search operators such as 'AND', 'OR', and 'NOT'. 'Free text searching' enables more informal or 'natural' search specifications, that is, either lists of search terms without search operators, or approximations to natural language (NL) expressions (e.g. questions). Of course, conventional search-engines do not analyze or exploit syntactic information in these NL expressions, and the grammatical terms expressing syntactic relations are ignored by the search engine. Search results are, to a first approximation, a ranking of indexed documents based on the relevance of those documents to the search specification. Relevance ranking methods vary, but, generally, the more search terms a document contains (and the more unique those terms are across the collection) the more relevant, and hence, the more highly ranked a document is in the 'hit list' [3].

This characterization of search glosses over many important variations in techniques for indexing, ranking and interpreting query specifications. For example, modern Web search methods also rank documents in terms of hyperlink patterns, independent of specific search specifications: important Web pages are likely to be those that have relatively numerous links to other pages, or are frequently linked from other pages [4]. Nonetheless, the characterization of search is close enough for present purposes.

Searching can be useful for identifying a set of documents whose content has some probability of containing facts and relationships relevant to the search intention, but text searching itself is a coarse-grained way to access this information. Unless users are skilled in crafting Boolean search specifications, text results typically return too many documents. For example, a text query against MEDLINE for documents about 'cell cycle' AND '*Saccharmoyces*' returned 4909 abstracts in 2000 [2]. Other capabilities provide additional ways to help users focus on documents relevant to their search intention. Common search functions include 'fuzzy' operators that search for canonical forms of search terms, abstracting out morphological differences owing to differences in aspects such as tense and plurals. Fuzzy searching increases recall, or the number of documents that contain search terms. Proximity operators can require search terms to occur within a certain window of text (e.g. a sentence). These functions are aimed at increasing the precision of search, that is, the probability of finding documents with relevant combinations of terms. Although these features help, further enhancements to search quality exist. These involve the use of IE techniques to extract automatically terms or keywords from document content, and the use of these terms to enhance query formulation, and analyze search result documents.

Enhanced search through text mining

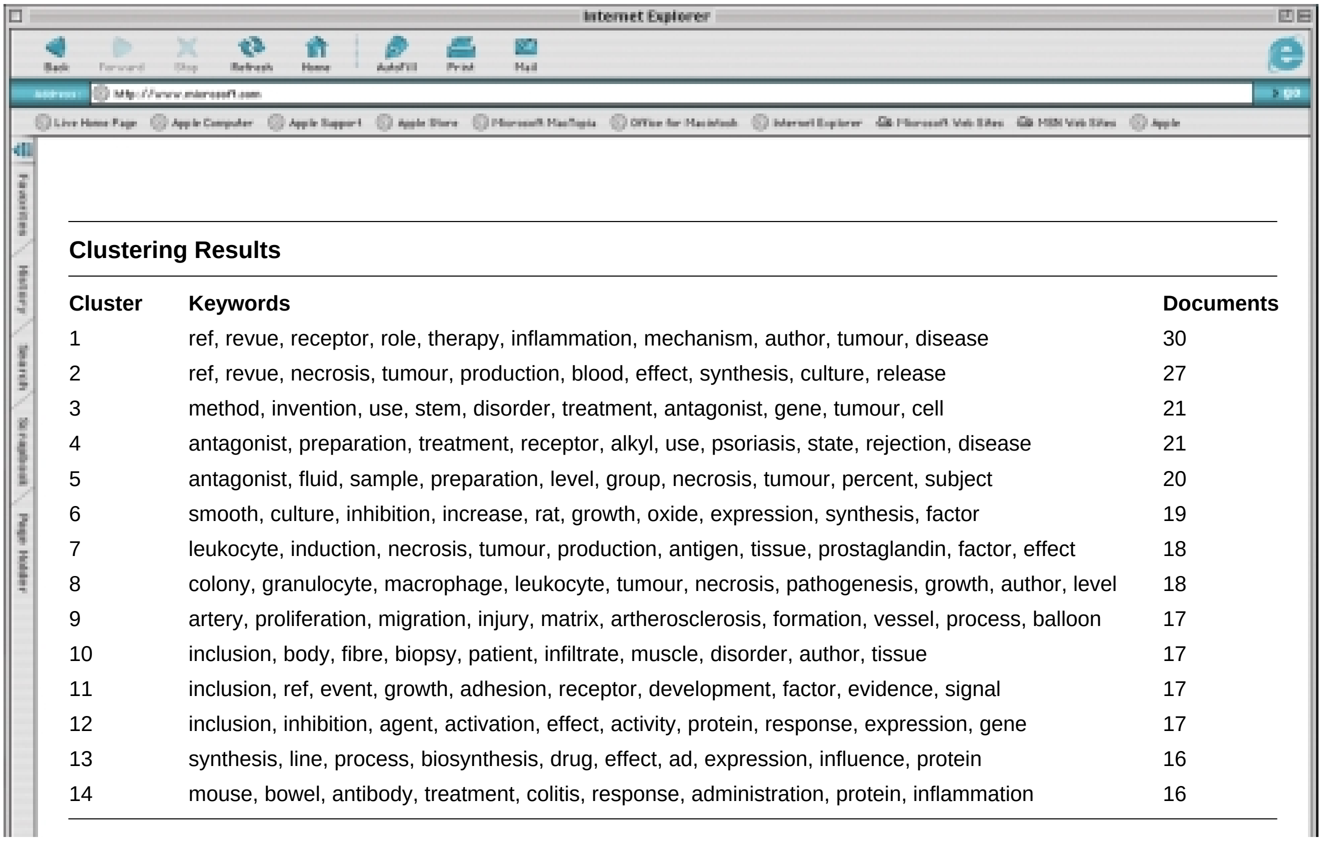
Several functions can help users focus on finding documents that answer the questions implied by their search specification. These enhancements are based on a more systematic analysis of document content, independent of the users' search behavior. These methods aim to organize search results into clusters or categories, summarize documents, or help searchers find effective search terms in the first place. Recent surveys of many of these techniques in the broader context of knowledge management can be found elsewhere [3,5,6].

Query refinement

A typical problem in text searching is that the query terms specified by users are not the 'best' terms for expressing the information they require because the terms do not match those expressed in the documents. Query refinement is intended to help users identify terms that are more likely to be expressed in documents. One approach for query refinement is based on relevance feedback, where searchers select search documents that appear relevant to their search goal, and then ask to retrieve 'more documents like this'. The search application then formulates and executes a new and expanded search specification, which is automatically generated by the system using terms selected from the specified 'relevant' documents. There also exist IE methods to find terms that co-occur with search terms in the document within some specified window, such as 'within a sentence'. These co-occurring terms can be presented as query prompts that searchers can use to refine and/or expand an initial search specification [7]. Query-expansion methods are discussed in more depth in [3].

Natural-language searching

Natural-language searching (sometimes referred to as 'semantic searching') refers to approaches that enable users to express queries in 'more-natural language' terms, for example, as explicit sentences or questions. In contrast to conventional search engines, which do not use the syntactic or semantic information available in natural-language expressions, natural-language searches attempt to use such information in the search process. That is, in addition to indexing 'terms' in documents, natural-language



Clustering Results

Cluster	Keywords	Documents
1	ref, revue, receptor, role, therapy, inflammation, mechanism, author, tumour, disease	30
2	ref, revue, necrosis, tumour, production, blood, effect, synthesis, culture, release	27
3	method, invention, use, stem, disorder, treatment, antagonist, gene, tumour, cell	21
4	antagonist, preparation, treatment, receptor, alkyl, use, psoriasis, state, rejection, disease	21
5	antagonist, fluid, sample, preparation, level, group, necrosis, tumour, percent, subject	20
6	smooth, culture, inhibition, increase, rat, growth, oxide, expression, synthesis, factor	19
7	leukocyte, induction, necrosis, tumour, production, antigen, tissue, prostaglandin, factor, effect	18
8	colony, granulocyte, macrophage, leukocyte, tumour, necrosis, pathogenesis, growth, author, level	18
9	artery, proliferation, migration, injury, matrix, artherosclerosis, formation, vessel, process, balloon	17
10	inclusion, body, fibre, biopsy, patient, infiltrate, muscle, disorder, author, tissue	17
11	inclusion, ref, event, growth, adhesion, receptor, development, factor, evidence, signal	17
12	inclusion, inhibition, agent, activation, effect, activity, protein, response, expression, gene	17
13	synthesis, line, process, biosynthesis, drug, effect, ad, expression, influence, protein	16
14	mouse, bowel, antibody, treatment, colitis, response, administration, protein, inflammation	16

Figure 1. Document clustering resulting from the text search specification ‘Asthma’, with the result documents clustered by topics. The search input field is not shown. Each document cluster topic is represented by a list of terms or keywords automatically extracted from the documents in each cluster. This figure is an abstraction of a search result clustering screen (i.e. some UI elements are not shown), part of a ‘Text Knowledge Miner’ web-based search software tool (described more fully by searching for ‘TKM’ on http://www.ibm.com).

search methods extract and index higher level semantic structures composed of terms, and relationships between terms. This can be done in different ways (for general discussion see [3]).

An example of a natural-language search involves enabling searchers to enter explicit, well-formed questions. In a question-answering prototype system developed by Prager *et al.* [8], a question like ‘What are the symptoms of formaldehyde use in humans?’ would be analyzed in semantic terms as follows: the ‘what’ question-form would be interpreted as a search for terms expressing cause and effect between the terms defined as ‘symptoms’, and the chemical formaldehyde for ‘humans’. The semantic categories implied by these terms are explicitly represented as annotations CAUSE, SYMPTOM, CHEMICAL, in analyzing the query, and in the search index itself. The search is based not only on finding documents where both content terms like ‘formaldehyde’ or ‘symptoms’ and semantic annotations, co-occur within a window of text. An example of a relevant document would be one that contains a passage such as ‘Formaldehyde [CHEMICAL], which is widely used in building materials and furnishings, can cause [CAUSE] nose and throat irritation [SYMPTOM], coughing [SYMPTOM], skin rashes [SYMPTOM], headaches [SYMPTOM]...’ and so on. (The annotations in square brackets are, of course, implicit, and not visible to searchers).

The text and linguistic analysis methods used for this purpose are common to the IE methods discussed later in this review. The structural relationships between terms can also be indexed and searched more explicitly and directly (as compared with looking for co-occurrences in the Q&A system mentioned earlier), and such an approach to natural-language searching has been developed by Baclawski *et al.* [9] (see also http://www.jarg.com). (Note that ‘semantic searching’ is an active area of research, and there are other vendors with such approaches.)

Clustering documents

Document clustering is a well-known and useful technique for organizing large document collections, including document collections resulting from text searches (for background see [3,10]). Documents in a cluster have ‘similar content’ defined by salient terms that are common to those documents. Iliopouos

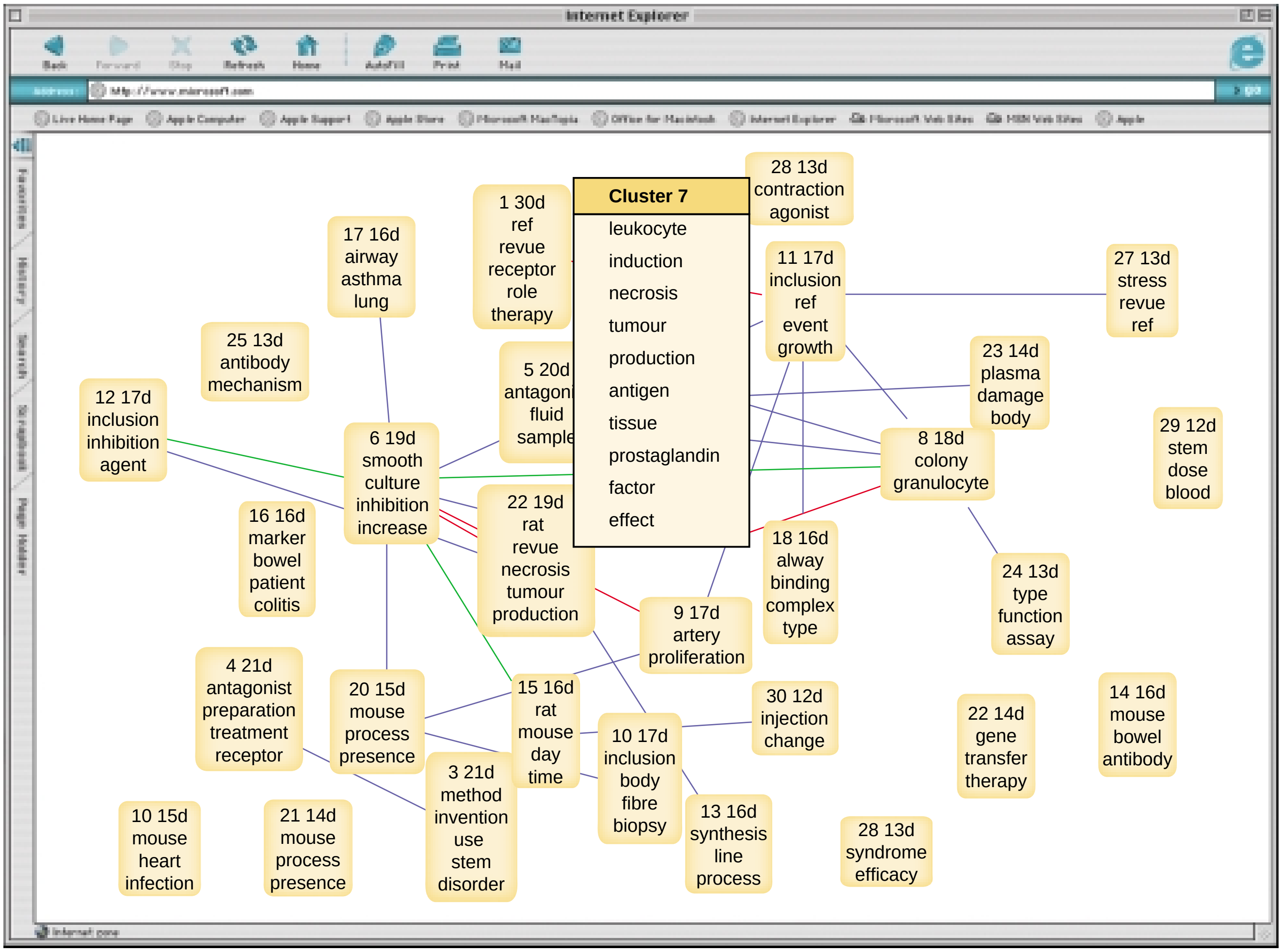

Internet Explorer
Back
Forward
Stop
Refresh
Home
AutoFill
Print
Mail
Address
http://www.microsoft.com
go
Cluster 7
leukocyte
induction
necrosis
tumour
production
antigen
tissue
prostaglandin
factor
effect
28 13d contraction agonist
1 30d ref revue receptor role therapy
17 16d airway asthma lung
11 17d inclusion ref event growth
27 13d stress revue ref
25 13d antibody mechanism
5 20d antagoni fluid sample
23 14d plasma damage body
12 17d inclusion inhibition agent
6 19d smooth culture inhibition increase
8 18d colony granulocyte
29 12d stem dose blood
16 16d marker bowel patient colitis
22 19d rat revue necrosis tumour production
18 16d alway binding complex type
24 13d type function assay
9 17d artery proliferation
4 21d antagonist preparation treatment receptor
20 15d mouse process presence
15 16d rat mouse day time
30 12d injection change
22 14d gene transfer therapy
14 16d mouse bowel antibody
10 17d inclusion body fibre biopsy
3 21d method invention use stem disorder
10 15d mouse heart infection
21 14d mouse process presence
13 16d synthesis line process
28 13d syndrome efficacy
Internet zone

Figure 2. Another view of documents retrieved by the keyword ‘Asthma’, and clustered using a graphic visualization of the document clusters. The lengths of the lines connecting the clusters are indicators of the strength of the semantic relationship of the clusters. The keywords shown in each cluster node are extracted from the clustered documents, and provide descriptions of the cluster topic. This figure is an abstraction of a search result clustering screen (i.e. some UI elements are not shown), part of a ‘Text Knowledge Miner’ web-based software tool (described more fully by searching for ‘TKM’ on http://www.ibm.com). The list of terms overlaid on the network (beginning with ‘leukocyte’) is a pop-up window expanding on the keywords describing the clusters that an end user has selected in the application.

et al. [11] developed an interesting variant to this technique by clustering documents by biological topics. They identified document clusters and the terms that best characterized those clusters by adapting IR measures of term frequencies, within and between document clusters. The defining terms are used to visually represent the cluster. These features can also be seen in a text search and text mining tool called Text Knowledge Miner (TKM; accessible via search from http://www.ibm.com), as shown in Figs 1 and 2. In both figures, salient terms extracted from the documents in each cluster are used as a description or label for the cluster. In the graphic view shown in Fig. 2, the links indicate the similarity of one cluster to another (in terms of the content of the documents in each cluster), and show the terms that label each cluster.

Categorizing documents

Another way to organize documents is through categorization. In categorization, categories and sub-categories collectively specifying a taxonomy are defined and named ahead of time by a domain expert. Several biotechnical knowledge resources (or ontologies) provide examples of taxonomies for classifying biological or chemical entities. The meaning of a category is defined by manually associating a set of representative documents (sometimes called training documents) with each category. Categorization tools also analyze the term content of these documents and create a representation of each category. Conventional categorization schemes are based on representation of text documents in an *n*-dimensional vector space defined by extracting terms from documents. Conversely, rule-based

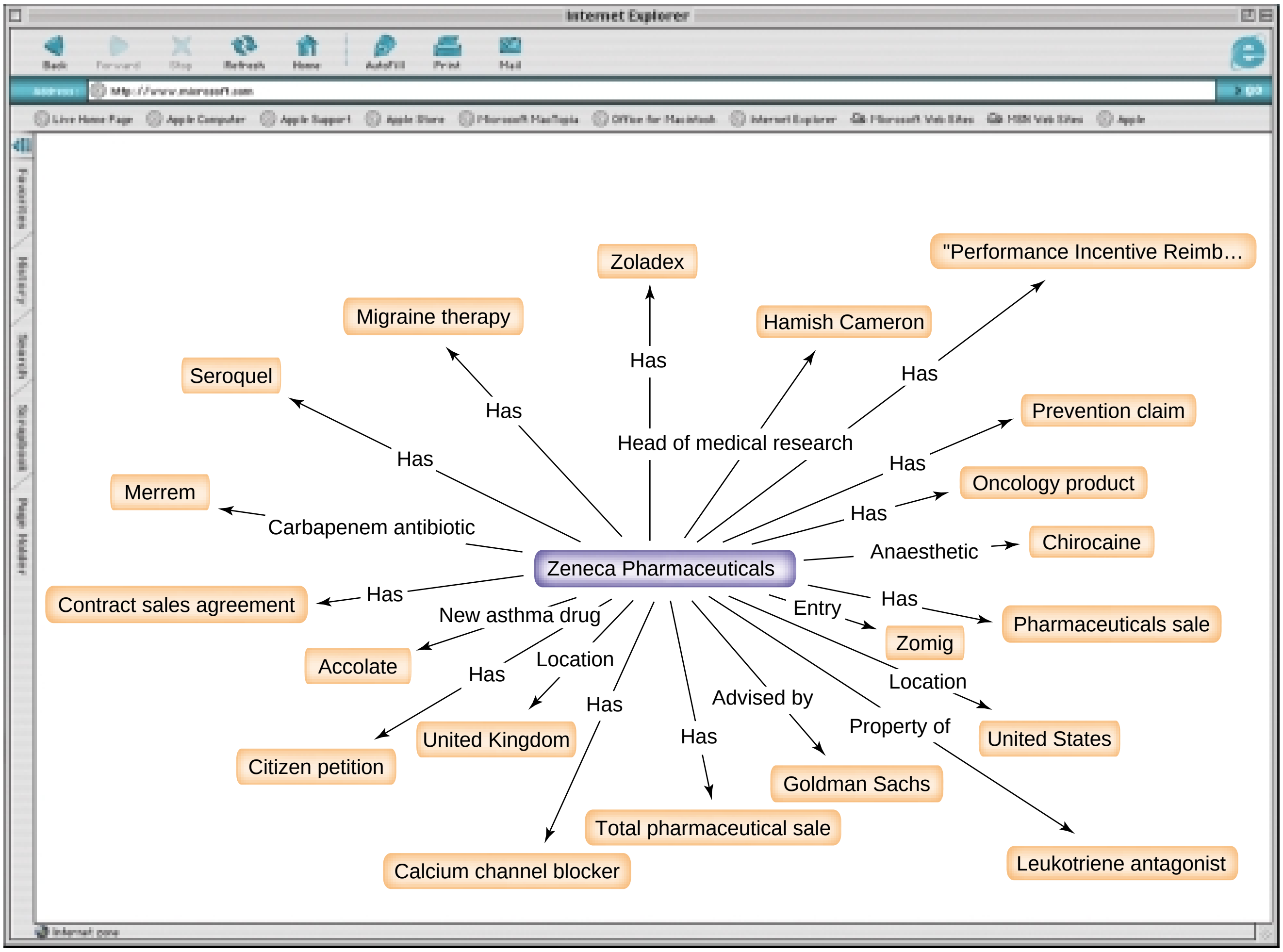


Figure 3. Lexical network for the root keyword 'Zeneca Pharmaceuticals', graphically representing terms related to the root keyword. The relationships have names extracted from the text context in which the terms co-occur, for example, 'property-of', 'location', and so on. This figure is an abstraction of an IBM research prototype (described more fully in [7]). 'Zeneca Pharmaceuticals' is a search term that is used to retrieve documents (in another part of the application), and is used to find related terms, that are represented to end users as a network of terms and links.

categorizers generate rules connecting terms and categories. These rules can be user-readable and -editable, illustrating how such tools assist humans in a difficult task, while still automating the key aspects of those tasks [3,12].

Summarizing documents

Summarization techniques are intended to generate automatically an abstract or summary of a full document that conveys the general idea of the document. An effective summary can help a searcher decide whether the document is relevant to their search without reading the full content. Summarization can be done at different levels. For example, a keyword summary collects terms extracted from a document and presents them as a summary. These terms can be collected as a separate 'summary text' shown in place of the full document. These terms can also be highlighted in the document content itself, to make it easier for searchers to quickly scan for these terms in the full document [13]. More sophisticated summarization techniques collect sentences that contain salient keywords relevant to a search specification or to the general topic of the document. The number of sentences, and hence the length of the summary, is a parameter that can be changed. The more sophisticated methods modify summary sentences selected from different parts of the document to form a more cohesive text, for example, by resolving pronoun references [14,15].

Information extraction

Information extraction methods automatically identify the 'entities' expressed in text as names, and the relationships between those names [16,17]. The enhanced text-search methods discussed previously are based, in part, on the extraction of terms (including names and phrases) from documents, which are then used by other text analysis methods such as document clustering, among other things. However, the promise of IE is more ambitious: extracting 'facts' at the level of specific propositions expressed in sentences and paragraphs (that is, a relationship between two or more named entities expressing a fact). This is an active area of bioinformatics research, and examples include: extraction of specific categories of biomedical terms such as gene or protein names [18,19]; relationships between these entities, such as protein–protein interactions [20]; terms describing functional attributes of proteins [21]; or relationships between genes, cell-lines and drug treatments [22].

Consider EDGAR, for example, a system developed by Rindflesch *et al.* for 'extracting drug, gene and relations [among them].' Like most IE tools, EDGAR uses a variety of linguistic methods to infer from sentences like (1), the semantic propositions or facts underlying them, as exemplified in (2) and (3):

(1) Compared with parental or mock-transfected HAG-1 cells, v-src-transfected HAG/src3-1 cells showed a 3.5-fold resistance to cisdiamminedicholoroplatinum (CDDP).

(2) Resistant (v-src, HAG/src3-1, CDDP).

(3) Cell ('99140404', 'HAG-1', 'gallbladder', 'adenocarinoma', tfw ('v- ... src'), 'human').

The proposition in (3) was constructed, in part, by associating the cell-line name 'HAG-1' with information derived from a knowledge source (or ontology) and annotating the extracted name from other facts known about the cell-line.

Visualizing terms and relationships graphically can be a useful way to interpret extracted information. An example is the 'lexical network' shown in Fig. 3. A lexical network is a graph showing terms and the relationships connecting them. These relationships can be named or unnamed. Figure 3 shows some simple entities and named relations based around the corporate entity '(Astra) Zeneca'. In this example, the relationship named 'location' is connected to the named country entity 'United Kingdom'. The lexical relations represented in Fig. 3 are generated using a variety of methods, including some level of linguistic parsing, and computation of statistical co-occurrence between terms [7]. Note that even if these methods cannot extract named relationships, unnamed relationships can be computed based on analyzing simpler co-occurrences between named entities – these relationships can also be useful.

The natural-language processing (NLP) methods necessary to analyze text at this level of meaning, vary in depth and sophistication, from statistical analysis methods defined over linguistic entities (that is, terms), to linguistic methods of varying depth and completeness, including syntactic parsing methods (for more discussion see [19,23,24]). At present, linguistic methods are still demanding in terms of linguistic expertise or the computational resources needed to apply these methods. The expertise required is that of writing grammars that can parse and identify the linguistic structures needed to extract 'entity' descriptions (e.g. noun phrase) and the relationships between them (e.g. expressed in verb phrases and syntax). Natural language is complex, typically expressing the same underlying semantic meaning in more than one surface forms of sentence expression. For example, 'protein A inhibits protein B' can be expressed in a variety of ways, such as active or passive voice, or as a variety of other descriptions (an interesting analysis can be found in [20]). IE methods try to extract the underlying meaning (fact) expressed by varied surface forms of textual discourse [25].

The linguistic methods used in IE are available in both research and commercial tools. For example, the tool shown in Fig. 3 is a research prototype based on an IBM 'Intelligent Miner for Text' product [7]. It should be noted here that several other commercial technologies exist for extracting terms and relations, in addition to tools for graphically representing entities and relations in various business and technical domains. Examples include: Inxight Software (http://www.inxight.com/.com), a spin-off of Xerox PARC Research; The Brain Technnologies Corporation (http://www.thebrain.com); and LexiQuest (http://www.lexiquest.com), recently acquired by SPSS (http://www.spss.com; Jouve, O. *et al.*, unpublished observations).

Semantic annotation and ontologies

One approach to recovering and representing the original semantic intentions expressed in the literature, is to have experts manually describe the key entities, their attributes and relationships between them, and use these descriptions to manually annotate (curate) the documents. In the biomedical domain, standards have been set for manual curating of biomedical documents. MEDLINE consists of abstracts annotated with a standard controlled vocabulary called Medical Source Headings or MeSH.

However, MeSH is more than a controlled vocabulary. It is also an example of an ontology, a framework of concepts and relationships expressed in terms, term relationships, synonyms, and categories that collectively express relatively stable knowledge about biomedical topics. Ontology relationships can be taxonomic, for example, classifying chemicals or drugs by various criteria, pharmacological or otherwise, or describing part-whole relationships between cells and cell structures. Other ontologies have been developed for other biomedical domains. The Unified Medical Language System (UMLS; sponsored by the US National Library of Medicine) is a comprehensive set of ontologies, a superset of specific ontologies, such as MeSH, among other ontologies (see http://www.nlm.nih.gov/research/umis).

Ontologies can be used effectively in information retrieval and information extraction. They can be used to help identify the 'named entities' in documents by mapping lexical terms extracted from documents to the corresponding terms in the ontology. MetaMap is a tool available with licensed access to UMLS that maps terms in documents to those in UMLS ontologies. Information extraction techniques have also been applied to annotate document content automatically using ontology terms. For example, Soumya *et al.* [26] (http://www.nlm.nih.gov/research/umis) have used categorization techniques to assign automatically Gene Ontology codes [27] expressing gene functions for genes referred to in MEDLINE abstracts. Baclawski *et al.* [9] have explored methods for annotating not only terms but also specific relationships expressed in text and ontologies (the authors call these term–relation structure keynets).

Of course, ontologies do not really solve the problem of extracting knowledge. Ontologies need to be updated, and although automatic annotation methods are promising, manual methods will be necessary for some time to come. Documents describing research findings or interpretations express potential new facts and implications that are as-yet not represented in an existing ontology. Building, extending and maintaining ontologies entails a host of difficult technical issues, including how to validate new facts and connections, and how to represent knowledge in a way to support inference and higher-level biological knowledge beyond atomic facts. Examples of research in the development of ontologies can be found elsewhere [28–30].

Knowledge discovery

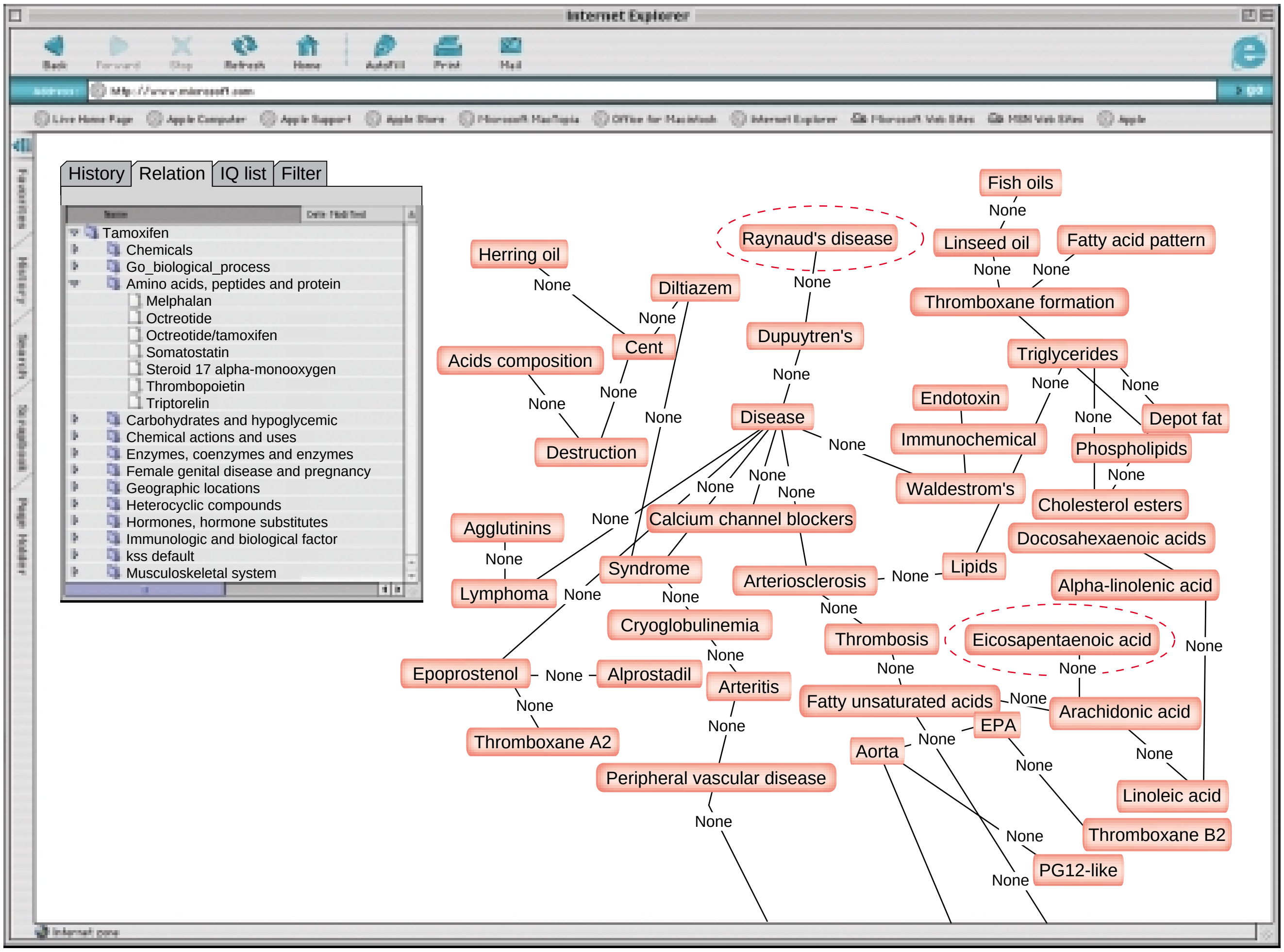
Information extraction potentially provides better support for text-based understanding than IR in that it extracts specific facts and implications originally intended by the writers of the text document. However, IE is ultimately a means to an end. Biomedical researchers need to discover new facts about genes and proteins and the biological contexts in which they function, with the aim of identifying new drug targets or disease treatments. Knowledge discovery methods involve compiling and integrating text descriptions and other kinds of data to create an interpretive context for understanding the meaning and implications of biological data. These contexts can range from informal models of terms and relationships, for example, term clusters or lexical graphs, to more formal models of biological structure and function, for example, metabolic pathways. These compilations not only focus on textual information, but also typically involve integration of information from multiple sources, including repositories of biological data, and organization of the compiled information into clusters, profiles or networks.

The MEDMINER prototype is an example of a system that searches and integrates information from text and data sources, and then analyzes and organizes the compiled information around topics that are relevant to the search specification [31]. MEDMINER searches PUBMED for MEDLINE documents, starting with a searcher's initial query specification (e.g. a combination of protein names and a relationship of interest, such as 'inhibit'), followed by an expanded text search automatically generated with additional biomolecular terms derived from a gene profiling tool called GeneCards. The latter tool searches biomolecular data sources, using the search entities (e.g. protein names) as search criteria, and then identifies name variants and synonyms, and uses these to create an expanded text query for additional searches. MEDMINER methods filter, organize and prioritize these expanded search results by looking for combinations of the search terms within the text. By contrast, searching PUBMED directly for documents containing these combinations of search terms would require searchers, following the authors' analysis, to create complicated Boolean search specifications.

The lexical networks described earlier for graphical visualization of terms and relationships, also provide a tool for discovering potential new connections among terms (Fig. 3). Lexical networks show relationships between terms, and these relationships can express either semantic relationship (e.g. 'located in') or an unidentified 'mutual co-occurrence' relationship between terms. Mutual co-occurrence measures do not result in identification of the relationship, but they do provide a connection, a strength of relationship, and links to the original document contexts containing the unnamed relationship.

An example of the potential use of unnamed relationships is to identify 'hidden links' among terms across collections of documents that do not directly (one-to-one) connect the terms. The concept of hidden links derives from work by Swanson, which started in the mid-eighties, and found relationships between disease syndromes and dietary or other chemical substances that could treat diseases [32,33]. For example, Swanson found a connection between Raynaud's disease and fatty acids in fish oil. The novelty of Swanson's work is that these relationships were not (at that time) actually expressed in specific sentences or documents. Instead, the disease syndrome was connected to an intermediate cluster of concepts in an intermediate literature collection associated with blood aggregation and viscosity.These concepts were, in turn, connected via an entirely different document context to the dietary or chemical substances in question.

Initially, Swanson used essentially manual methods and clever inference to deduce this relationship (he later developed specialized computer tools for supporting these analyses). However, an alternative lexical network tool shown in Fig. 4 shows the connection directly as it emerged from searching for named and unnamed relationships in the same two sets of literature orginally used by Swanson. Of course, lexical networks can be quite complicated, and additional tools are required to manage the networks. In Fig. 4, the taxonomy tree on the left of the screen corresponds to the biomedical MeSH taxonomy. The lexical network terms are mapped into this taxonomy, and then



- Tamoxifen
  - Chemicals
  - Go_biological_process
  - Amino acids, peptides and protein
    - Melphalan
    - Octreotide
    - Octreotide/tamoxifen
    - Somatostatin
    - Steroid 17 alpha-monooxygen
    - Thrombopoietin
    - Triptorelin
  - Carbohydrates and hypoglycemic
  - Chemical actions and uses
  - Enzymes, coenzymes and enzymes
  - Female genital disease and pregnancy
  - Geographic locations
  - Heterocyclic compounds
  - Hormones, hormone substitutes
  - Immunologic and biological factor
  - kss default
  - Musculoskeletal system

Figure 4. An alternative version of the lexical network shown in Fig. 3. This version exemplifies the 'hidden links' between 'Raynaud's disease' (node at the top), and fatty acids in fish oil (e.g. 'fish oils' at the top or 'eicosapentaenoic acid' in the lower right). This figure is an abstraction of an IBM research prototype (described more fully in [7]). The links represent relationships between terms, and 'none' means these relationships are not yet identified. The indented tree on the left is a set of MeSH categories, and extracted terms in MEDLINE documents that are classified under MeSH categories. For example, 'Melaphalan' and 'Octreotide' are terms that appeared in MEDLINE abstracts, which are instances of the MeSH category 'Amino acids, peptides, and proteins'. These MeSH categories can be used to filter the number and categories of terms that are depicted in the network.

the taxonomy categories are used to filter out entire classes of terms thereby focussing on subsets of terms and relationships, and increasing the likelihood of finding useful connections.

Several studies, in somewhat different contexts, demonstrate the value of integrating text and biomolecular data into networks or clusters of terms, annotated with various categories of terms. For example, Jenssen *et al.* [34] identified co-occurrences of gene names within the scope of an entire MEDLINE abstract. They constructed from these names a graph of gene co-occurrences, assigning strengths to the relationships based on frequency of co-occurrence (although they do not appear to use the specific mutual co-occurrence measure described previously). Next, they used the gene names in the network to search for other literature references involving the genes, and annotated the network with these descriptions. These compiled annotations, in the context of the network relationships among genes, suggested real connections to the corresponding gene products and their role in various cell processes of disease.

Stapley and Benoit [2] also compiled networks of genes based on co-occurrence of gene references in text (they used a 'bibliometric' measure of co-occurrence strength). They used the genes named in these networks to search for and compile biomolecular data on the genes. Kankar *et al.* [35] analyzed gene clusters starting from biomolecular experiments. Gene clusters derived from microarray experiments were used to search for MEDLINE documents referring to these genes. The researchers analyzed the distribution of MeSH keywords used to annotate the abstracts across genes, and then clustered these keywords into general topics about the entire gene cluster, and subtopics about subsets of the genes. Both studies exemplify the integration and compilation of text and biomolecular data with the aim of creating an interpretive context for the genes of interest.

Andrade and Valencia [21] focussed on compiling information about proteins. They used term frequency to annotate proteins described in MEDLINE documents with terms describing the functions of proteins. Terms were selected for annotation when they occurred more frequently with the protein family to which a target protein belonged, compared with frequency of occurrence with other protein families. Comparisons to manually annotated abstracts suggest the validity of this method.

Text mining can also be used to enhance data mining. Chang *et al.* demonstrated how text literature can improve the likelihood of finding valid homologies or similarities between human gene sequences and those of other species [36]. Similarities in gene sequences could imply similarities in gene products and the role they play in cell and organism functioning, hence providing a basis for making inferences about human gene sequences from information known about simpler organisms. However, these inferences are not always valid. Chang *et al.* showed that, in some cases, valid homologies also correlate with text descriptions of the genes involved in the sequences. That is, combining data-based homology searches with text searches of collateral information can help refine the homology search process, thereby increasing the precision of finding relevant and real homologies.

Discovery can also refer to the compilation of smaller facts into larger representations that express more-complicated biological structures and functions – for example, metabolic pathways in cell function. These can be represented as complex interconnected networks of proteins and enzymes that interact in cell processes. Many sources of data and experimentation are needed to infer such networks, but one of these sources can be text literature. Ng and Wong, for example, have demonstrated that protein–protein interactions extracted from text can be compiled to infer sub-networks of larger pathways [1].

In all these examples, the relationships between terms were all expressed in the text documents, and, theoretically, are available for researchers to read, remember and comprehend. The overwhelming issue is that this human-mediated understanding is increasingly difficult, given the huge volume, complexity and specialization of the literature. Text-mining techniques can help humans represent the content of large collections of documents in ways that make the implications easier to comprehend.

Frontiers of text-based knowledge discovery

There is great opportunity for improving the text-mining methods outlined in this review. A key research focus is discovering more-effective methods for inferring facts and implications, and relating them to more complex and formal models of biological structure and function. Current IE methods focus on relatively small facts, but, of course, text documents express more complicated concepts as well. These could involve propositions expressed in multiple sentences, in more-complicated discourse structures and arguments, or even in multiple documents and non-text data sources. Advances in IE, natural language processing, integrated text and data mining, and knowledge representation schemes, all hold the promise of automating human-like capabilities for comprehending complicated knowledge structures. A good source of leading-edge research in these areas include the Pacific Symposium on Biocomputing (PSB), which began in 1996 (http://psb.stanford.edu); worth noting are the online tutorials with annotated biobliographies, presented at PSB 2001 by J-I. Tusujii and S. Ananiadou (HTTP://www-tsujii.is.s.u-tokyo.ac.jp/ ~genia/ tutorial/) [37].

The development and application of ontologies will continue to be an active area of research. Ontologies reflect the fact that small facts are themselves connected to form larger entities, expressing what we know about a topic. Developing ontologies is not simple, and will require innovation in both how knowledge is represented and organized in larger structures than propositions, and how these structures can be used to make inferences. We believe that ontologies will be important for capturing, representing and managing knowledge as it emerges from different sources, including text mining. As ontologies evolve into more-complex knowledge representations, the relationship between ontologies and biological models will need to be explored (see [28,29]).

Finally, linguistic structures could do more than just express knowledge. An emerging research direction is potential language-based approaches to biological modeling and prediction. This hypothesis is based on the observation that – like language – biological entities (e.g. protein molecules) are made up of a hierarchy of structures, and that these structures and their interactions might be captured in grammar-like structures and rules (Institute for Research in Cognitive Science; http://www.ircs.upenn.edu/modeling2000) [38].

References

1 Ng, S-K. and Wong, M. (1999) Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Infomratics. 1, 104–112*

2 Stapley, B. and Benoit, G. (2000) Biobliometrics: information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. *Proc. Pac. Symp. Biocomput. 5*, 529–540

3 Baeza-Yates, R. and Ribeiro-Neto, B., eds (1999) *Modern Information Retrieval*. ACM Press, New York

4 Chakrabarti, D. *et al.* (1999) Mining the Web's link structure. *Computer* 32, 60–67

5 Mack, R. *et al.* (2001) Knowledge portals and the emerging digital knowledge workplace. *IBM System Journal* 40, 925–955

6 Marwick, A. (2001) Knowledge management technology. *IBM System Journal* 40, 814–830

7 Cooper, J. and Byrd, R. (1997) Lexical navigation: visually prompted query expansion and refinement. *Proc. 2nd ACM Conf. Digital libraries*, 23–26 July 1997, Philadelphia, PA, USA, pp. 237–246

8 Prager, J. *et al.* (2000) Question-answering by predictive annotation. *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2000, Athens, Greece, pp. 184–191

9 Baclawski, K. *et al.* (2000) Knowledge representation and indexing using the unified medical language system. *Proc. Pac. Symp. Biocomput*. 5, 493–504

10 Cutting, D. *et al.* (1992) Scatter/Gather: a cluster-based approach to browsing large document collections. *Proc. 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, June 1992, Copenhagen, Denmark, pp. 318–329

11 Iliopouos, I. *et al.* (2001) TEXTQUEST: document clustering of MEDLINE abstracts for concept discovery in molecular biology. *Pac. Symp. Biocomput. 6*, 384–395

12 Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys* (in press)

13 Neff, M. and Cooper, J. (1999) ASHRAM: active summarization and markup. *Proc. Annual HICSS Conference*, 5–8 January 1999, Maui, Hawaii, USA, p. 83

14 Ando, R.Y. *et al.* (2000) Multidocument summarization by visualizing topic content. *Proc. ANLP/NAACL Workshop on Automatic Summarization*, pp. 79–88

15 Boguraev, B. and Neff, M. (2000) Lexical Cohesion, discourse segmentation and document summarization. *Proc. RIAO*, 12–14 April 2000, Paris, France

16 Ravin, Y. and Wachholder, N. (1996) Extracting names from natural language text. *IBM Research Report* (http://www.research.ibm.com/people/r/ravin)

17 Vilvaldi, J. and Rodriquez, H. (2000) Improving term extraction by combining different techniques. In *Workshop on Computational Terminology for Medical and Biological Applications* (Ananiadou, S. and Maynard, D., eds), pp. 61–68, Springer-Verlag

18 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

19 Fukuda, K. *et al.* (1998) Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput. 3*, 707–718

20 Blaschke, C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pp. 60–67

21 Andrade, M. and Valencia, A. (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts: development of a prototype system. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5, 25–32

22 Rindflesch, T. *et al.* (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.*5, 538–549

23 Nasukawa, T. and Nagano, T. (2001) Text analysis and knowledge mining. *IBM Systems Journal* 4, 967–984

24 Nobata, C. *et al.* (1999) Automatic term identification and classification in biology texts. *Proc. of the Natural Language Pacific Rim Symposium*, pp. 369–375

25 Yakushijii, A. *et al.* (2001) Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput. 6*, 408–419

26 Soumya, R. *et al.* (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.* 12, 203–214

27 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology: the Gene Ontology Consortium. *Nat. Genet.* 25, 25–29

28 Rzhetsky, A. *et al.* (2000) A knowledge model for analysis and simulation of regulatory networks in bioinformatics studies aiming at disease gene discovery. *Bioinformatics* 16, 1120–1128

29 Yu, H. *et al.* (1999) Representing genomic knowledge in the UMLS semantic network. *Proc of AMIA Symposium*, pp. 181–185

30 Hahn, U. *et al.* (2002) Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pac. Symp. Biocomput. 7*, 338–349

31 Tanabe, L. *et al.* (1999) MedMiner: an internet text-mininig tool for biomedical information, with application gene expression profiling. *Biotechniques*. 27, 1210–1214, 1216–1217

32 Swanson, D. (1999) Implicit text linkages between MEDLINE records: using Arrowsmith as an aid to scientific discovery. *Library Trends* 48, 48–59

33 Weeber, M. *et al.* (2001) Using concepts in literature-based discovery: simulating Swanson's raynaud-fish oil and migraine-magnesium discoveries. *J. Am. Soc. Inf. Sci. Tech.* 52, 548–557

34 Jenssen, T-K. *et al.* (2001) A literature network of human genes for high throughput analysis of gene expression. *Nat. Genet.* 28, 21–28

35 Pankar, P. *et al.* MedMeSH Summarizer: text mining for gene clusters. *Proc SIAM Conference in Data Mining, SDM 2002 (in press)*

36 Chang, J. *et al.* (2001) Including biological literature improves homology search. *Proc. Pacific Symposium on Biocomputing*, ??–?? Month Year, Location, pp. 374–383

37 Tusujii, J-I. and Ananiadou, S. (2001) Tutorial: information extraction from scientific texts. *Pacific Symposium on Biocomputing 2001* (http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/tutorial/)

38 IRCS and Center for Bioinformatics (2001) *Workshop on Language Modeling of Biological Data*, 25–27 February 2001, Philadelphia, PA, USA (http://www.ircs.upenn.edu/modeling2000/)