A text-mining system for knowledge discovery from biomedical documents

by N. Uramoto
H. Matsuzawa
T. Nagano
A. Murakami
H. Takeuchi
K. Takeda

This paper describes the application of IBM TAKMI® for Biomedical Documents to facilitate knowledge discovery from the very large text databases characteristic of life science and healthcare applications. This set of tools, designated MedTAKMI, is an extension of the TAKMI (Text Analysis and Knowledge MIning) system originally developed for text mining in customer-relationship-management applications. MedTAKMI dynamically and interactively mines a collection of documents to obtain characteristic features within them. By using multifaceted mining of these documents together with biomedically motivated categories for term extraction and a series of drill-down queries, users can obtain knowledge about a specific topic after seeing only a few key documents. In addition, the use of natural language techniques makes it possible to extract deeper relationships among biomedical concepts. The MedTAKMI system is capable of mining the entire MEDLINE® database of 11 million biomedical journal abstracts. It is currently running at a customer site.

The life science industry is an emerging market in which application spaces, such as drug discovery and development in the pharmaceutical sector and clinical record management in health care, have become areas of significant recent interest.[1] Documents in the scientific literature play an important role in life science by serving as a potential source for underlying knowledge discovery. These documents are a rich repository of information on relationships among biomedical concepts such as genes, proteins, diseases, and a variety of other key topics.

Text mining is a technology that makes it possible to discover patterns and trends semiautomatically from huge collections of unstructured text.[2–6] It is based on technologies such as natural language processing, information retrieval, information extraction, and data mining.[7] Early papers in this area mentioned the possibility of knowledge discovery from the biomedical literature. Hearst, one of the founders of text mining, proposed a system for predicting the functions of unknown genes using biomedical documents.[2] Swanson also described the idea of discovering new knowledge from the biomedical literature.[4] Subsequently, considerable research has been done in the areas of biomedical concept extraction (named-entity extraction), relationship extraction, and network/pathway construction for protein-protein interaction. However, although text mining has proved a promising approach for knowledge discovery from text sources, certain specific problems are encountered when trying to apply it to the realm of life science.

First, existing approaches are incapable of handling the vast amount of textual domain-specific information available. Indeed, there is more data available than anyone could possibly read or digest. For ex-

0018-8670/04/$5.00 © 2004 IBM

ample, MEDLINE**[8] is a database of over 11 million citations (abstracts) of biomedical articles dating back to the 1960s. MEDLINE is widely used as a golden standard for text-mining systems in life science, and several text-mining applications using MEDLINE have been proposed. The MedMeSH Summarizer[9] extracts MeSH** (Medical Subject Headings) terms[10] that can summarize the nature of a cluster of gene names obtained from DNA microarrays (also called DNA chips). MedMiner[11] is a system that filters information for the PubMed** search engine.[12] Obviously, any approach that applies text-mining methods to such a large document collection must be highly scalable and robust.

Second, existing information extraction systems only provide extracted concepts and relationships in a fixed way. Because these systems are noninteractive, it is difficult to iteratively apply mining processes on their results directly. With an interactive text-mining system, users are better able to discover hidden knowledge by using a combination of mining functions and a trial-and-error approach.

To address these problems we have developed a text-mining system called IBM TAKMI* for Biomedical Documents (designated MedTAKMI hereafter), which is capable of mining the entire MEDLINE database in an interactive manner. The predecessor of this system, TAKMI (Text Analysis and Knowledge MIning), is a text-mining system for customer relationship management (CRM), which has been successfully used in call centers to mine customer support call logs.[13] The MedTAKMI system extends TAKMI to provide a useful set of tools for knowledge discovery from biomedical documents. MedTAKMI is designed to handle large document sets and is thus capable of mining the entire set of MEDLINE citations.

The development of methods for extracting information on such biomedical concepts as genes, proteins, and diseases from text is an active area of research[14–19] and typically involves the following two primary subtasks:

1. *Entity extraction*—the recognition of gene, protein, and chemical names from biomedical text
2. *Relation extraction*—the extraction of relationships among these entities

Thus architecturally MedTAKMI consists of two main components designed to handle information extraction and entity/relationship mining.

The MedTAKMI system performs entity extraction based on dictionary lookup. This approach is simple conceptually and can recognize entities very quickly. We have developed a large domain dictionary that contains two million biomedical entities. These entities and their associated category names are used as keywords in the MedTAKMI system so that users can search for documents that contain a keyword within a specific category, for example, a query on the keyword "p53" within the gene category.

In a preprocessing stage input documents are parsed by a shallow syntactic parser which extracts keywords (entities) with category labels, as well as any binary and ternary relationships that may exist among these entities. The MedTAKMI runtime engine then uses this information to provide mining functions to users. Categories are constructed from public ontological knowledge, for example, using the MeSH terms in MEDLINE or the resources provided by Gene Ontology**.[3] User-defined resources may also be employed.

There has been extensive research in relation extraction,[20–35] wherein the goal is to extract relationships among biomedical entities (e.g. proteins and genes), from patterns such as "A inhibits B" and "A activates B," where A and B represent specific entities. Such relationships may be extracted by using one or more of the following information and methods:

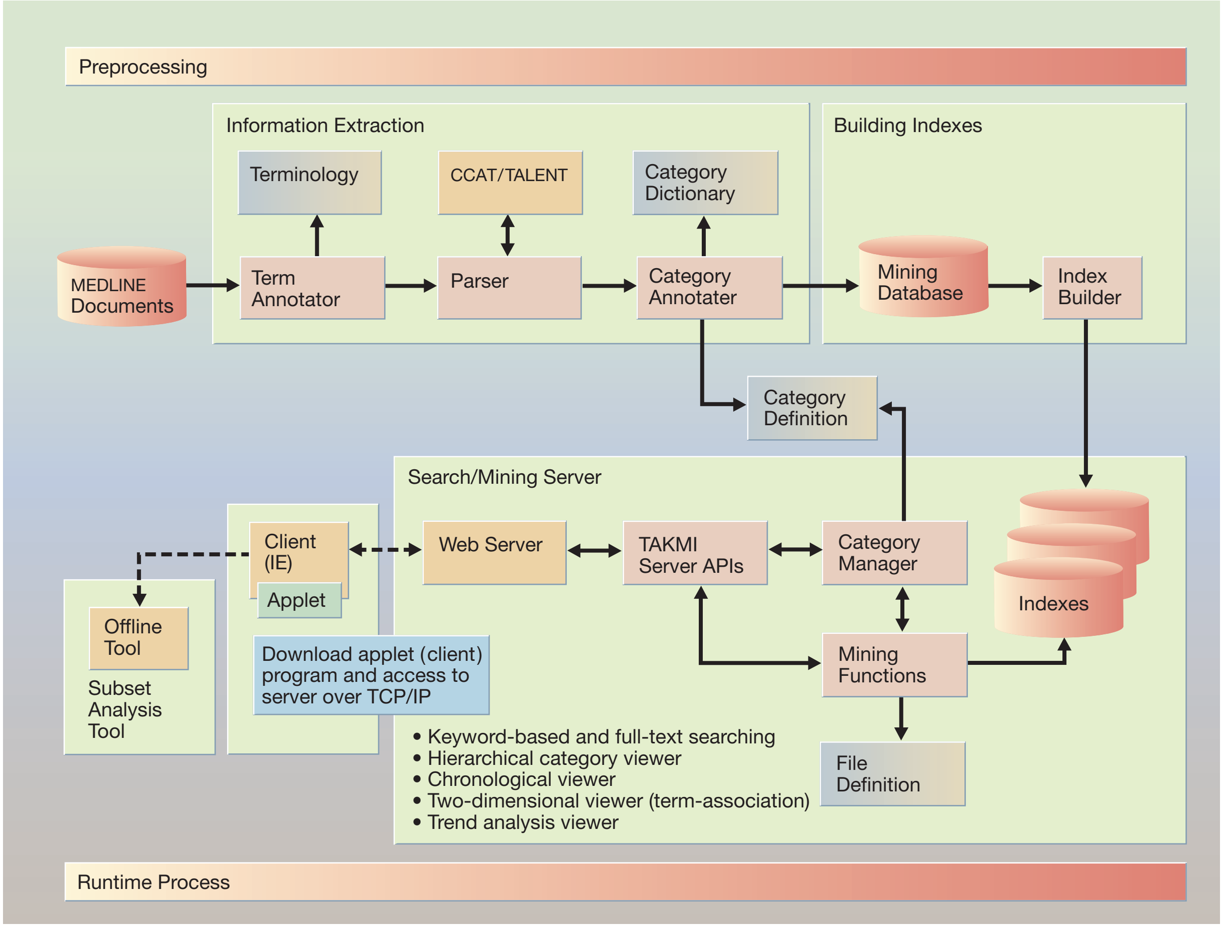
- Surface string patterns[20]
- Syntactic information from shallow parsing[21,22] and full parsing[23–27]
- Templates and rules[28–30]
- Statistical information with machine learning[32–35]

In particular, the MedTAKMI system uses syntactic information with a shallow parser to extract binary (a noun and a verb) and ternary (two nouns and a verb) relationships. These relationships from the document collection are aggregated and can be displayed by category viewers as described later.

As previously noted, MedTAKMI is an extension of TAKMI, a text-mining system for CRM.[13] The main differences between these two systems are the following:

- The use of hierarchical categories: TAKMI only supports flat categories such as product names. For MedTAKMI we developed a hierarchical category viewer because most biomedical entities (e.g., genes and diseases) are defined hierarchically.

Figure 1 MedTAKMI architecture



- The extraction of ternary relationships to capture protein-protein interaction by using deeper language analysis
- The introduction of support for domain-specific mining functions
- The development of a new system architecture and componentization structure

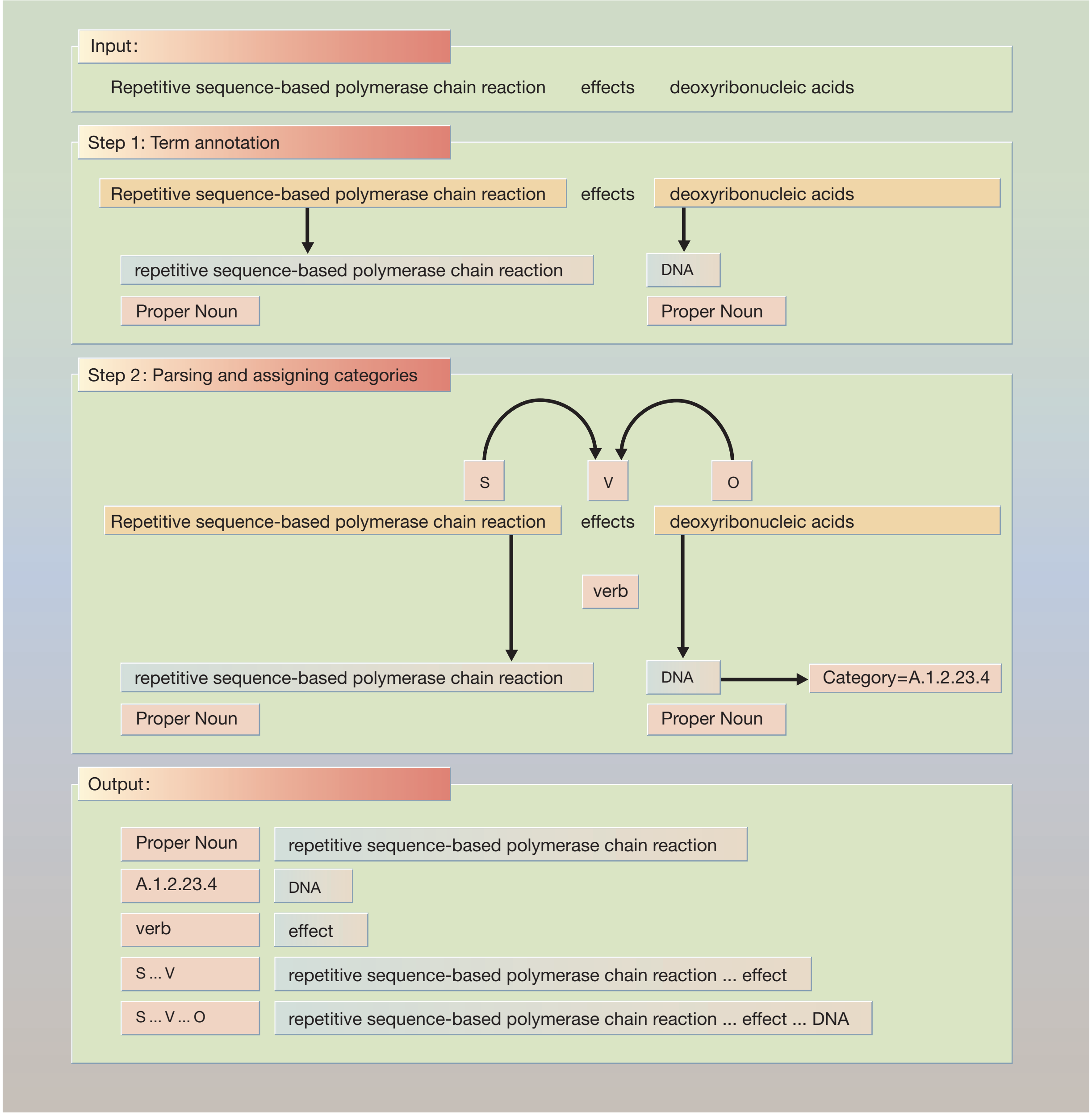
This paper is organized as follows: The next section describes the key features of MedTAKMI, including the system architecture and the information extraction process. We then introduce the searching and mining functionalities of MedTAKMI. The following section provides an example of the application of the MedTAKMI system to a specific user scenario. Finally, we summarize our work and describe directions for future research.

Features of the MedTAKMI mining system

The MedTAKMI architecture consists of two main components: a preprocessing information extraction stage and a runtime search/mining server, as shown in Figure 1. In this section we briefly discuss these main components and then describe methods for information extraction.

Information extraction process. Information extraction occurs as a preprocessing step and involves several subcomponents (see Figure 1, upper left). In Step 1, the term annotator finds words in the input text (i.e., the sentence with the label “Input” shown in Figure 2) using the term dictionary and identifies these words using their canonical form. (As described in more detail later, the term dictionary contains a

Figure 2 Information extraction process


Input:
Repetitive sequence-based polymerase chain reaction
effects
deoxyribonucleic acids
Step 1: Term annotation
Repetitive sequence-based polymerase chain reaction
effects
deoxyribonucleic acids
repetitive sequence-based polymerase chain reaction
DNA
Proper Noun
Proper Noun
Step 2: Parsing and assigning categories
S
V
O
Repetitive sequence-based polymerase chain reaction
effects
deoxyribonucleic acids
verb
repetitive sequence-based polymerase chain reaction
DNA
Category=A.1.2.23.4
Proper Noun
Proper Noun
Output:
Proper Noun
repetitive sequence-based polymerase chain reaction
A.1.2.23.4
DNA
verb
effect
S ... V
repetitive sequence-based polymerase chain reaction ... effect
S ... V ... O
repetitive sequence-based polymerase chain reaction ... effect ... DNA

pair of forms for each term: a surface form and a canonical form.) The obtained canonical words are embedded in the text document as annotations in XML (eXtensible Markup Language). In Step 2, the annotated text is passed to a syntactic parser. The parser outputs segments of phrases labeled with their syntactic roles, for example NP (noun phrase) or VG (verb group). The category annotator then assigns categories to the terms in these segments and phrases. The category dictionary consists of a set of canonical forms and their categories, which also indicates the node label in the hierarchy of categories. The hierarchical categories are in turn imported from existing hierarchies, such as the MeSH terms in MED-

LINE, or from user-designed resources. Syntactic relationships among these entities, for example, subject-verb (S-V) or subject-verb-object (S-V-O), are also extracted from the output of the parser. All extracted information is finally encoded into an index file that is used by the runtime part of the system.

Search/mining server process. The search and mining server, shown in the lower part of Figure 1, provides users with the searching and mining services described in detail later. The MedTAKMI system is a Web application and its client code is loaded into Internet Explorer** as an applet (a servlet version is also available now). Note that hierarchical category definitions are shared by both the information-extraction and the runtime-server preprocessing components.

MEDLINE—a collection of biomedical documents. In their work life-science researchers typically use MEDLINE,[8] a bibliography database that covers the biomedical area. To understand our approach to the extraction of information from MEDLINE, some understanding of MEDLINE itself is required. MEDLINE is administered by the National Center for Biotechnology Information (NCBI)[36] of the United States National Library of Medicine (NLM).[37] It contains approximately 11 million biomedical citations, dating from the mid 1960s to the present. Citations in MEDLINE are collected from over 4600 biomedical journals published worldwide. Biomedical citations in MEDLINE are available to the general public at the PubMed Web site.[12] Figure 3 shows an example of one such citation.

To make citation lookup easy, NLM indexes the articles in MEDLINE for retrieval and classification using the Medical Subject Headings (MeSH) thesaurus.[10] MeSH headings consist of sets of descriptors in a hierarchical structure. Subheadings, or qualifiers, provide additional specificity for each descriptor. The major MeSH headings indicate the main contents of the article, and the minor MeSH headings are used to describe secondary topics. MeSH also contains other features such as check tags and age tags.

Each citation contains the article title, abstract, authors' names, MeSH headings, affiliations, publication date, journal name, and other information. The title and abstract are text strings that can be manipulated by natural-language-processing text-mining techniques. For example, the MEDLINE citation shown in Figure 3 contains the information shown in Table 1.

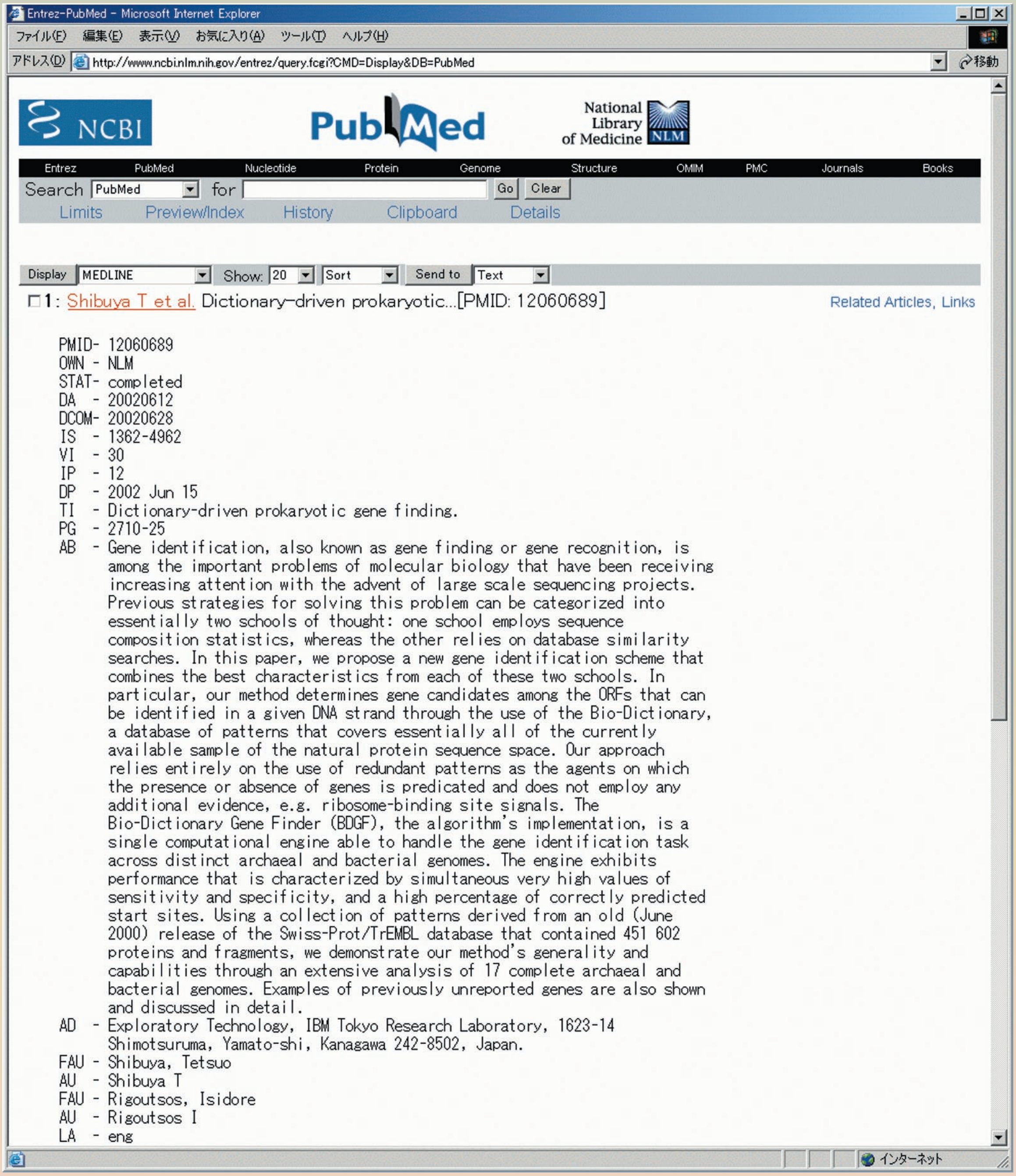
Figure 4 shows the result of retrieving the MeSH descriptor "Amino Acid Sequence" using the MeSH Browser.[38] These results indicate that there are two nodes labeled "Amino Acid Sequence" in the descriptor thesaurus: one at G06.184.603.060 and the other at L01.453.245.667.060. These codes represent locations; for example, G06.184.603.060 is the child node of Molecular Structure (G06.184.603). The letter *G* at the beginning of the location represents the major category Biological Sciences, while *L* represents another major category, namely, Information Science.

Dictionary-based information extraction. The preprocessing stage thus extracts meta-data information from MEDLINE documents. In addition to information such as author and publication date, the system can extract information from other text fields, for example, title or abstract, using natural language processing. This information can include sets of keywords (e.g. protein names), predicate-argument binary relations (e.g. "activate-protein"), and ternary relations (e.g. subject-verb-object dependency triplets).

In the preprocessing phase of the MedTAKMI system, document titles and abstracts are parsed by CCAT, a shallow syntactic parser developed at the IBM Thomas J. Watson Research Center using an approach originally proposed by Charniak.[39] Because this is a general-purpose parser that has not been trained for biomedical documents, it is difficult to obtain optimized results by parsing documents from the medical domain. We solve this problem by first annotating the text with domain dictionaries. The term dictionaries are constructed by users and employ resources such as UMLS** (Unified Medical Language System)[40] or the users' own proprietary resources. The annotations facilitate the parsing of medical-domain text even when the parser has not been specifically trained for this domain. This annotation process is needed for two primary reasons:

1. *Identification of term boundaries*—Most technical terms in the medical domain, for example, protein names, are compound words. Thus, biomedical terms tend to consist of a combination of numerals, symbols, and verbs, making it very difficult to find term boundaries. For instance, the compound noun "repetitive sequence-based polymerase chain reaction" consists of an adjective (repetitive), a past participle of a verb (se-

Figure 3 PubMed Web site


Entrez-PubMed - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Display&DB=PubMed
移動
NCBI
PubMed
National Library of Medicine NLM
Entrez PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search PubMed for
Go Clear
Limits Preview/Index History Clipboard Details
Display MEDLINE Show: 20 Sort Send to Text
Related Articles, Links
インターネット

□1: Shibuya T et al. Dictionary-driven prokaryotic...[PMID: 12060689]

```
PMID- 12060689
OWN - NLM
STAT- completed
DA  - 20020612
DCOM- 20020628
IS  - 1362-4962
VI  - 30
IP  - 12
DP  - 2002 Jun 15
TI  - Dictionary-driven prokaryotic gene finding.
PG  - 2710-25
AB  - Gene identification, also known as gene finding or gene recognition, is
      among the important problems of molecular biology that have been receiving
      increasing attention with the advent of large scale sequencing projects.
      Previous strategies for solving this problem can be categorized into
      essentially two schools of thought: one school employs sequence
      composition statistics, whereas the other relies on database similarity
      searches. In this paper, we propose a new gene identification scheme that
      combines the best characteristics from each of these two schools. In
      particular, our method determines gene candidates among the ORFs that can
      be identified in a given DNA strand through the use of the Bio-Dictionary,
      a database of patterns that covers essentially all of the currently
      available sample of the natural protein sequence space. Our approach
      relies entirely on the use of redundant patterns as the agents on which
      the presence or absence of genes is predicated and does not employ any
      additional evidence, e.g. ribosome-binding site signals. The
      Bio-Dictionary Gene Finder (BDGF), the algorithm's implementation, is a
      single computational engine able to handle the gene identification task
      across distinct archaeal and bacterial genomes. The engine exhibits
      performance that is characterized by simultaneous very high values of
      sensitivity and specificity, and a high percentage of correctly predicted
      start sites. Using a collection of patterns derived from an old (June
      2000) release of the Swiss-Prot/TrEMBL database that contained 451 602
      proteins and fragments, we demonstrate our method's generality and
      capabilities through an extensive analysis of 17 complete archaeal and
      bacterial genomes. Examples of previously unreported genes are also shown
      and discussed in detail.
AD  - Exploratory Technology, IBM Tokyo Research Laboratory, 1623-14
      Shimotsuruma, Yamato-shi, Kanagawa 242-8502, Japan.
FAU - Shibuya, Tetsuo
AU  - Shibuya T
FAU - Rigoutsos, Isidore
AU  - Rigoutsos I
LA  - eng
```

Table 1 Meta-data for a MEDLINE citation

Symbol	Meaning	Value
PMID	PubMed Identifier	12060689
TIS	ISSN for the journal	1362-4962
VI	Volume	30
DP	Published Date	2002 Jun 15
TI	Title	Dictionary-driven prokaryotic gene finding
AB	Abstract	Gene identification, also known as gene finding or gene recognition, is among the important problems of molecular biology that have been receiving increasing attention with the advent of large scale sequencing projects.......... (Snipped.)
AD	Affiliation	Exploratory Technology, IBM Tokyo Research Laboratory, 1623-14 Shimotsuruma, Yamato-shi, Kanagawa 242-8502, Japan.
AU	Author	Shibuya, Tetsuo
AU	Author	Rigoutsos, Isidore
MH	MeSH heading	Algorithms
MH	MeSH heading	Amino Acid Sequence
MH	MeSH heading	Base Sequence
MH	MeSH heading	Codon, Initiator
MH	MeSH heading	Computational Biology/*methods (* represents Major MeSH)
MH	MeSH heading	*Genes, Archaeal
TA	Journal Title Abbreviation	Nucleic Acids Res

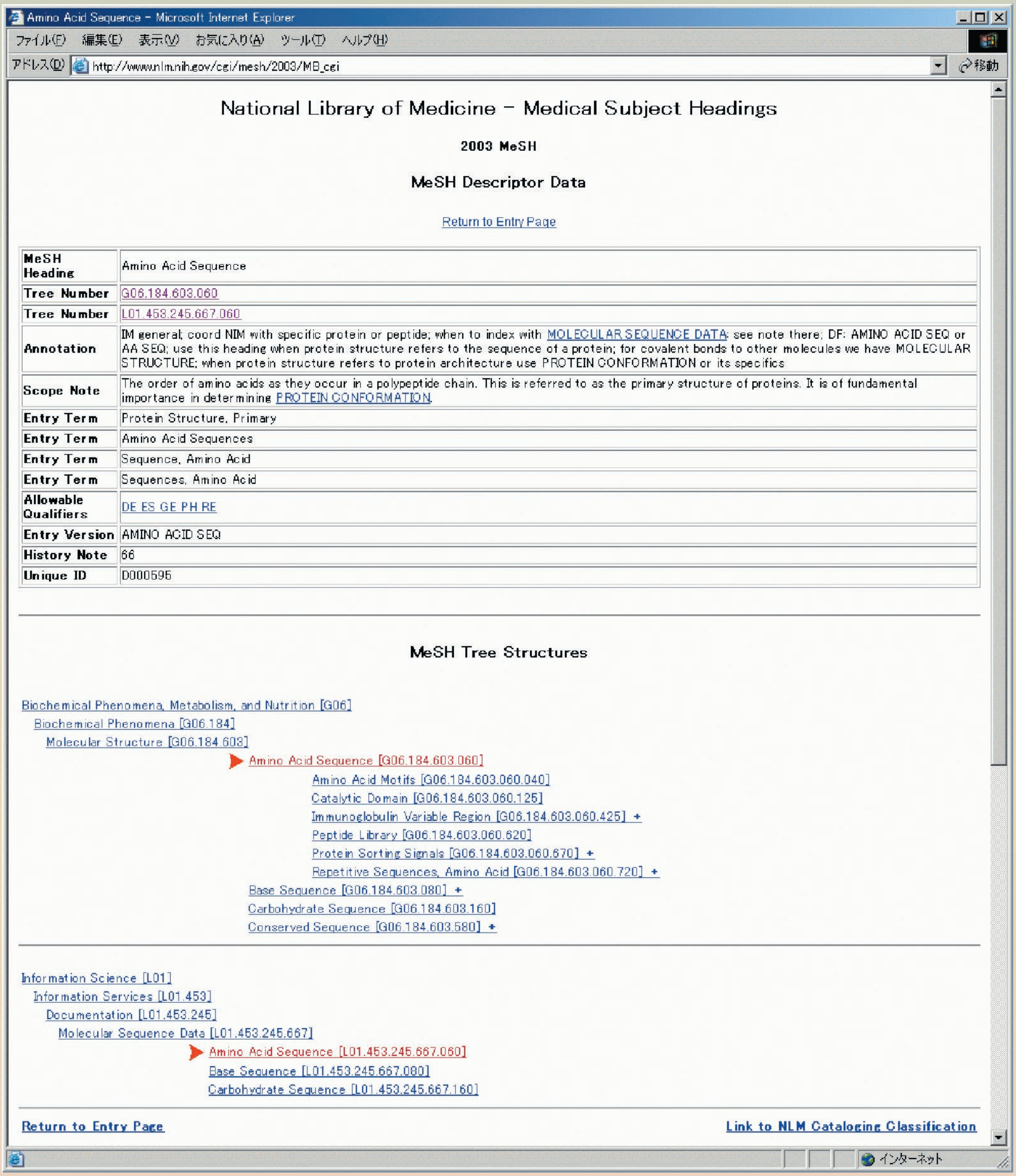
quence-based) and three nouns (polymerase, chain, reaction). The annotations made by the technical term dictionary are based upon a part-of-speech (POS) analysis. Note that words, such as *chain,* which can be a noun or a verb, can further complicate this process.

2. *Aggregation of synonymous expressions and spelling variations*—There can be multiple expressions that are synonymous with a particular technical term. These can arise from abbreviations or acronyms as well as from spelling variations. If these variations are recognized as different entities, it can often cause problems for text mining. For instance, "DNA" and "deoxyribonucleic acid" are synonyms; thus, they should be counted as the same entity in the mining process. The dictionary contains spelling/abbreviation variants and their canonical forms. By reducing these variants to a single canonical form, we can treat them as the same entity.

After text is annotated with a technical term dictionary it is parsed by a parser, or tagger. In this system, we use the CCAT parser to assign a POS to each word. This determination is based on the statistical distribution of candidate POSs for a word and the probability of POS transitions (from/to adjoining words) that are extracted from a training corpus. After the annotation and parsing processes, phrases are determined by using head-driven models,[41,42] and category identifiers are assigned. The current MedTAKMI system maintains approximately 270000 hierarchical category identifiers. Of these, roughly 170000 are MeSH category identifiers; the rest are user-defined categories. MeSH terms are updated annually, and the revisions are incorporated into MedTAKMI.

Relation extraction. In general, a conventional text analysis system counts the frequency of occurrence of a given term and calculates its importance. The

Figure 4 MeSH browser


Amino Acid Sequence - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://www.nlm.nih.gov/cgi/mesh/2003/MB_cgi
移動
インターネット

National Library of Medicine - Medical Subject Headings

2003 MeSH

MeSH Descriptor Data

Return to Entry Page

MeSH Heading	Amino Acid Sequence
Tree Number	G06.184.603.060
Tree Number	L01.453.245.667.060
Annotation	IM general; coord NIM with specific protein or peptide; when to index with MOLECULAR SEQUENCE DATA: see note there; DF: AMINO ACID SEQ or AA SEQ; use this heading when protein structure refers to the sequence of a protein; for covalent bonds to other molecules we have MOLECULAR STRUCTURE; when protein structure refers to protein architecture use PROTEIN CONFORMATION or its specifics
Scope Note	The order of amino acids as they occur in a polypeptide chain. This is referred to as the primary structure of proteins. It is of fundamental importance in determining PROTEIN CONFORMATION.
Entry Term	Protein Structure, Primary
Entry Term	Amino Acid Sequences
Entry Term	Sequence, Amino Acid
Entry Term	Sequences, Amino Acid
Allowable Qualifiers	DE ES GE PH RE
Entry Version	AMINO ACID SEQ
History Note	66
Unique ID	D000595

MeSH Tree Structures

- Biochemical Phenomena, Metabolism, and Nutrition [G06]
  - Biochemical Phenomena [G06.184]
    - Molecular Structure [G06.184.603]
      - ► Amino Acid Sequence [G06.184.603.060]
        - Amino Acid Motifs [G06.184.603.060.040]
        - Catalytic Domain [G06.184.603.060.125]
        - Immunoglobulin Variable Region [G06.184.603.060.425] +
        - Peptide Library [G06.184.603.060.620]
        - Protein Sorting Signals [G06.184.603.060.670] +
        - Repetitive Sequences, Amino Acid [G06.184.603.060.720] +
      - Base Sequence [G06.184.603.080] +
      - Carbohydrate Sequence [G06.184.603.160]
      - Conserved Sequence [G06.184.603.580] +

- Information Science [L01]
  - Information Services [L01.453]
    - Documentation [L01.453.245]
      - Molecular Sequence Data [L01.453.245.667]
        - ► Amino Acid Sequence [L01.453.245.667.060]
        - Base Sequence [L01.453.245.667.080]
        - Carbohydrate Sequence [L01.453.245.667.160]

Return to Entry Page

Link to NLM Cataloging Classification

Table 2 Examples of ternary relationships extracted from MEDLINE

apoptosis. . .induce. . .p53
apoptosis. . .inhibit. . .tumor protein p53 (Li-Fraumeni syndrome)
tumor protein p53. . .play. . .role
adhesion receptors. . .activate. . .extracellular signal-regulated kinase
Ang II. . .activated. . .p38

system can then find documents that contain the query terms or their synonyms. It may also display relationships among various terms using visualization tools. This would allow us to discover terms that are strongly related, but we still would not know how or in what way they are related. For example, we could find all documents containing the word *smoking*. But even though *smoking* may appear frequently within these documents, conventional text analysis systems are unable to tell us, for example, what the effects of smoking are, who is affected, and to what extent. We could next identify terms that are strongly related to the word *smoking*, but simple co-occurrence of two terms (e.g., *smoking* and *risk*) reveals very little about the semantic relationship between them. Conventional text analysis systems ignore hints that indicate whether a verb is negative or not and whether it is active or passive. Such systems are unable to distinguish between the meanings of the following two sentences: "Smoking increases the risk of lung cancer," and, "The risk of lung cancer increases the consequences of smoking." Both sentences contain the terms *smoking* and *risk*; simple co-occurrence detection is not sufficient to derive relationships.

The TAKMI system for CRM[13] is able to extract "subject . . . verb" or "verb . . . object" relationships in a sentence. It can extract such concepts as "modem . . . broken," "file . . . not found," and "hard disk . . . slow" from customer queries. These extractions were instrumental in facilitating problem detection and workload reduction for analysts at customer help centers. Although the binary "subject . . . verb" or "verb . . . object" relationships suffice for customer support purposes, a more detailed concept extraction is needed for biomedical articles. This is especially true because the "what subject influences what object" relationship is a very important one in the biomedical domain. For instance, it is clear that extraction of the relationship "protein_A. . .activates . . .protein_B. . .with. . .enzyme_C" provides more detailed information than either "protein_A. . .activates. . .protein_B" or "protein_B. . .with. . .enzyme_C." Furthermore, we need to distinguish between "protein_A. . .activates. . .protein_B. . .with. . . enzyme_C" and "protein_B. . .activates. . .protein_A. . .with. . .enzyme_C." In the MedTAKMI system, we extract these ternary relationships (as well as binary relationships) using natural-language-processing technology. These extracted relationships are used as keywords by various MedTAKMI mining functions, as described in the next section. Table 2 shows actual examples of ternary relationships extracted from MEDLINE documents.

Searching and mining functions

The MedTAKMI system is a text-mining system that can be integrated with search engines. It supports keyword-based search engines and can also be used with other search engine implementations such as IBM's GTR (Global Text Retrieval) and DB2* Net Search Extender. The mining process can be applied to the whole database or to a subset document collection obtained by a series of searches. Users can submit a query and receive a document collection in which each document contains the query keywords or their synonyms. Mining functions can then be applied to the collection in order to discover underlying information, such as protein-protein relationships. Alternatively, users can continue the search process by using the results of previous searching and mining operations. Interactivity is a very important MedTAKMI feature because it allows users to switch between searching and mining in a flexible manner.

MedTAKMI provides various mining functions for large document collections. Some of these functions are general, but others are tailored to the life-science domain. The following functions are introduced in this section:

- Keyword-based and full-text searching
- Hierarchical category viewer
- Chronological viewer
- Two-dimensional viewer (term-association)
- Trend analysis viewer
- Other analytical tools

Most of these viewers can function interactively, a unique feature of the MedTAKMI system. Experiences with text mining in the CRM domain[13] have previously demonstrated the importance of interactivity in that application, and interactivity also proves to be an important requirement for users in the life-science domain. For example, in knowledge-discovery and data-mining (KDD) tasks, such as on-line analytical processing (OLAP), users may wish to use a fact discovered in a previous mining result to structure a new query from a different point of view. Indeed, a user often cannot define a complete query beforehand but rather must iteratively refine the query based on previous results. Furthermore, when a document collection is very large, a single query may not be able to sufficiently narrow the collection, requiring a user to submit a sequence of queries to obtain the desired subset of documents. MedTAKMI's viewers help users to navigate toward this goal interactively.

Keyword-based and full-text searching. The MedTAKMI system provides two types of searching: keyword and full text. A keyword index which associates keywords with category codes is built by the information extraction phase, as described above. Users can thus submit a query such as "search for documents that contain the word p53 as a gene name." A disadvantage of keyword search, however, involves inaccuracies in keyword extraction. If a term consisting of multiple words is not recognized as a keyword in the indexing phase, then this term cannot be matched later in a query submitted by a user. Thus, in MedTAKMI, users can also choose a full-text search which uses different indexes built separately to retrieve documents in which the term appears literally. The MedTAKMI system allows users to switch between these two search engines seamlessly.

Hierarchical category viewer. The hierarchical category viewer shows keyword distribution in a data collection over a predefined hierarchy. For example, Figure 5 shows the distribution of keywords for the "Disease" node and its child nodes in the MeSH hierarchy.

Blue bars at the bottom of the viewer show the frequency of occurrence of a keyword. In Figure 5, there are 35 documents that contain the term "neoplasms" or any of the terms in its child nodes. (Note that for efficiency reasons this frequency was calculated for a subset of documents sampled from the full document collection. The cardinality of such a subset can be specified by the user.)

Red bars indicate relative frequency—this measure compares the current document subcollection to the initial document collection indexed in the MedTAKMI system. In other words, searching is a process that narrows down a document collection. Assume D is the initial document collection. A keyword search due to query q_1 returns D_1, the subset of D that satisfies q_1. Similarly, let $(q_1, q_2, \ldots, q_n)$ be a sequence of successive queries and let $(D_1, D_2, \ldots, \mathrm{D}_n)$ be document collections such that D_i is the result of query q_i made also on collection D. The relative frequency for a keyword w in the document collection D_i is calculated using the following formula:

$$relfreq(w, D_i) = \frac{\dfrac{c(w, D_i)}{|D_i|}}{\dfrac{c(w, D)}{|D|}}$$

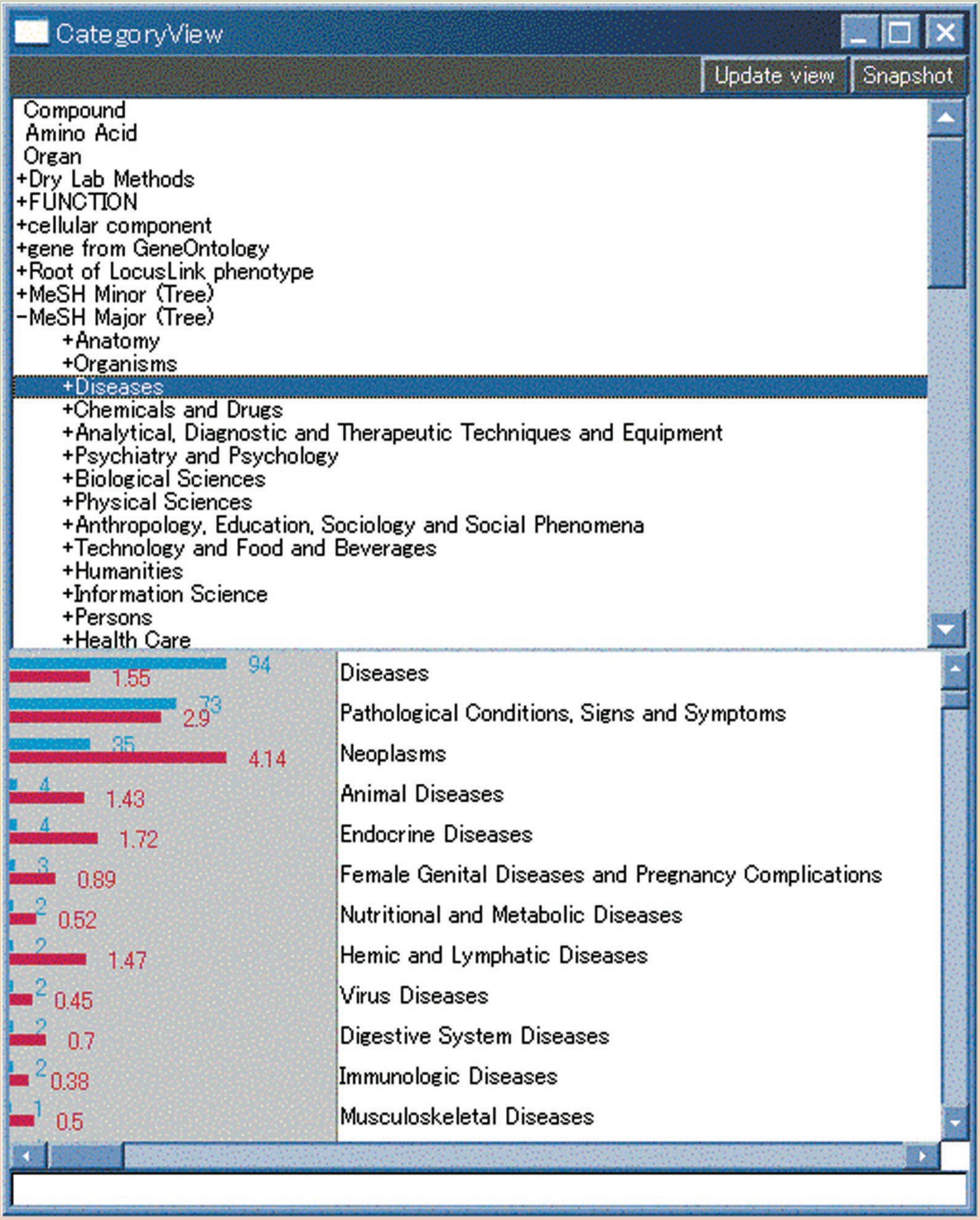
where $|D_i|$ is the number of documents in the document collection D_i and $c(w, D_i)$ is the number of documents that contain the word w in the collection D_i.

Two types of hierarchies are registered with the MedTAKMI system. The flat type has no children; for example, in Figure 5, "Compound," "Amino Acid," and "Organ" are flat. The tree type appears with a '+' notation; thus "Dry Lab Methods" and "MeSH Minor" in Figure 5 are tree type. Users can define the hierarchy set using public hierarchies such as MeSH and Gene Ontology[3] as well as user-specific or company-specific hierarchies. We have developed a hierarchy that currently consists of 95 000 nodes for the various concepts in the life-science domain.

Chronological viewer. This viewer allows a user to discover trends by viewing the chronological distribution of a set of documents. Figure 6 shows the monthly distribution of approximately 330 000 documents selected from MEDLINE in 2002. MedTAKMI supports yearly, monthly, and daily distributions. Using this viewer, one can determine when a certain keyword began to appear in the biomedical literature and how its frequency of occurrence changed with time. For example, the term HIV does not appear in MEDLINE documents in the 1970s, but by the 1990s it occurs with high frequency.

Two-dimensional maps (term-association). The two-dimensional viewer allows a user to visualize the strength of association between keywords. Figure 7 shows protein-protein associations in a mouse document collection. The value in each cell represents the strength of association of two keywords—the higher the value, the stronger the association. For example, the proteins "Bcl2-associated X protein"

Figure 5 Hierarchical category view



Compound
Amino Acid
Organ
+Dry Lab Methods
+FUNCTION
+cellular component
+gene from GeneOntology
+Root of LocusLink phenotype
+MeSH Minor (Tree)
−MeSH Major (Tree)
+Anatomy
+Organisms
+Diseases
+Chemicals and Drugs
+Analytical, Diagnostic and Therapeutic Techniques and Equipment
+Psychiatry and Psychology
+Biological Sciences
+Physical Sciences
+Anthropology, Education, Sociology and Social Phenomena
+Technology and Food and Beverages
+Humanities
+Information Science
+Persons
+Health Care

94	1.55	Diseases
73	2.9	Pathological Conditions, Signs and Symptoms
35	4.14	Neoplasms
4	1.43	Animal Diseases
4	1.72	Endocrine Diseases
3	0.89	Female Genital Diseases and Pregnancy Complications
2	0.52	Nutritional and Metabolic Diseases
2	1.47	Hemic and Lymphatic Diseases
2	0.45	Virus Diseases
2	0.7	Digestive System Diseases
2	0.38	Immunologic Diseases
1	0.5	Musculoskeletal Diseases

and “G elongation factor” have a strong association. The numbers “19 (36.54%)” in the cell mean that there were 19 documents which mentioned both Bcl2-associated X protein and G elongation factor, and that these 19 documents comprise 36.54 percent of the documents mentioning Bc12-associated X protein. Because 36.54 percent is much higher than other percentages, the cell is automatically highlighted.

Formally, the value for each cell $v(wx_i, wy_j, D)$ is calculated by using the following formula:

$$p(wx_i, wy_j, D) = \frac{c(wx_i \cap wy_j, D)}{c(wy_j, D)}$$

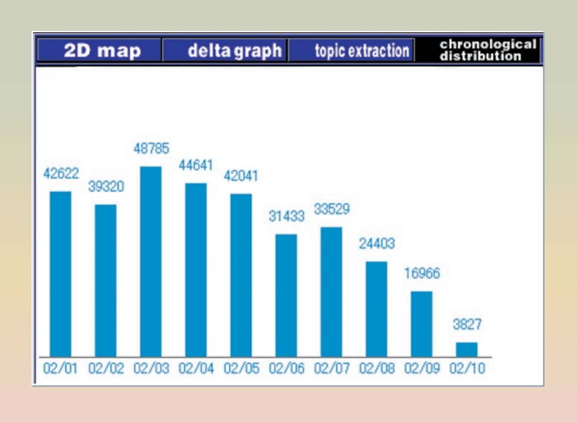
$$v(wx_i, wy_j, D) = \frac{p(wx_i, wy_j, D)}{\frac{1}{M}\sum_{j=1}^{N} p(wx_i, wy_j, D)}.$$

where $c(w, D)$ is the number of documents that contains the word w in the collection D. The words wx_i and wy_j belong to the categories of x and y axes, respectively. M and N are the size of the matrix for each x and y axis. Clicking a cell leads to further searching using the two keywords.

Users are not limited to protein-protein interactions. Associations for protein-gene, gene-disease, disease-organ, and so forth, are also possible. In fact, any combination of hierarchy terms is possible.

Other analytical tools. The viewers described thus far are interactive, allowing for a flexible mining process. However, some tools require a much longer analysis time to produce a response. These tools are available to the user as off-line tools in the MedTAKMI system. During the mining process intermediate results can be stored and used as input to an off-line tool. The final output of the off-line tool is shown independent of the runtime MedTAKMI process. This section introduces two off-line functions provided by MedTAKMI.

Figure 6 Chronological view



Similar entry search tool. The first function is one that permits searching of similar biological entities. This function uses the term-association table data described in the previous section. Each row of the table is represented by a vector whose attributes correspond to the keywords in the columns and whose elements represent the joint occurrence probabilities of the two corresponding entities. The degree of similarity between row entities is calculated based on the distance between their corresponding vectors using a cosine measure. Principal component analysis is used to reduce dimensions and improve precision because the number of columns tends to be large. Figure 8 shows the ranking result of signaling proteins similar to protein 14_3_3.

Subset analysis tool. The second function allows for the category view data or term association view data of two different queries to be compared and analyzed to identify significant differences. Consider two data sets A and B. The occurrence probabilities of entries that appear in both data set A and data set B are compared. Entries are first organized into the following three categories based on the AIC (Akaike Selection Criterion) approach for statistical model selection criteria.[43]

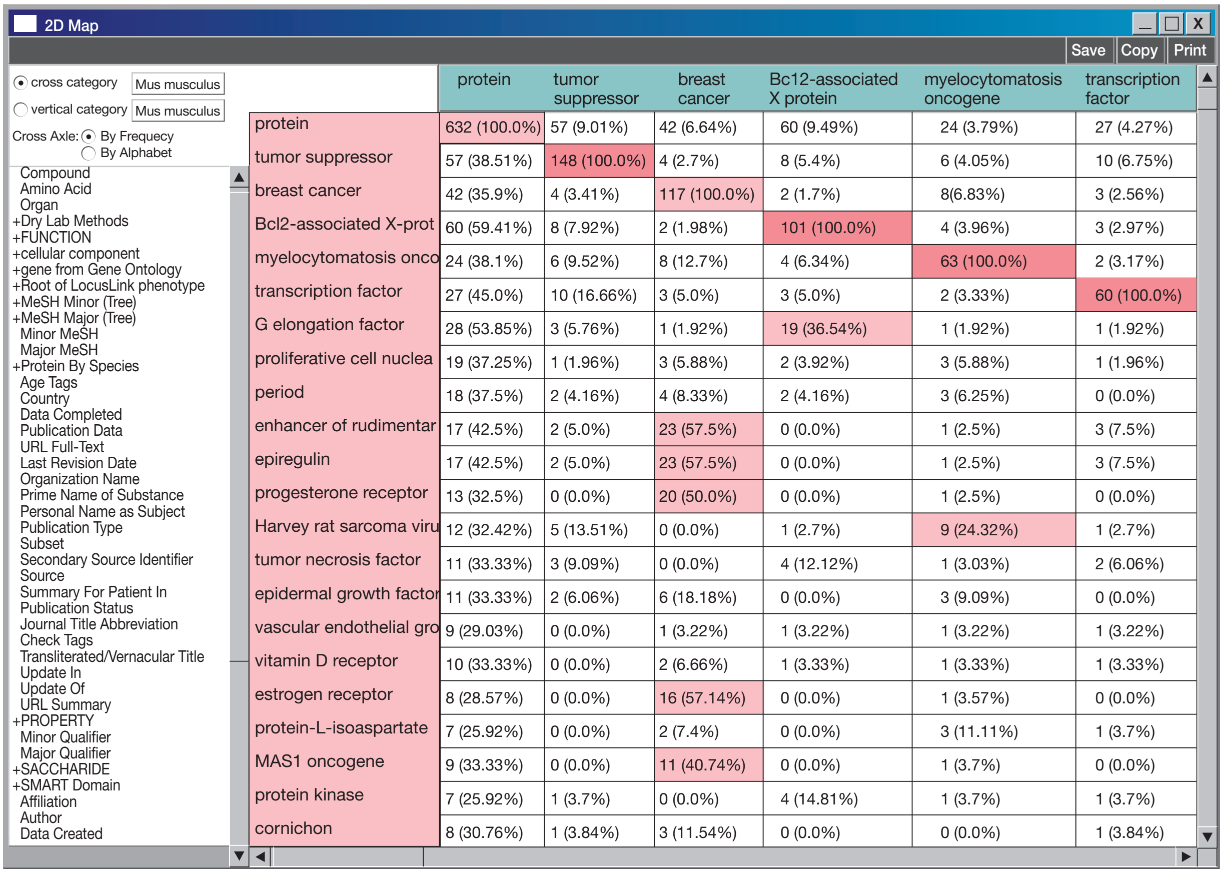
- Entries frequently appearing in data set A
- Entries frequently appearing in data set B
- Entries common to both data set A and data set B

For each category, entries are ranked by a score based on a statistical test. We use the following statistical test:

$$H_0: p_A = p_B$$

where p_A and p_B are the probability of occurrence of the entry in data set A and data set B respectively. The chi-square statistic is calculated from the occurrence data. Under the null hypothesis H_0, this statistic has a chi-square distribution with one degree of freedom χ_1^2. We define the following ranking score,

Figure 7 Two-dimensional view of proteins vs proteins (mouse-related) for a document set that contains the term “p53”

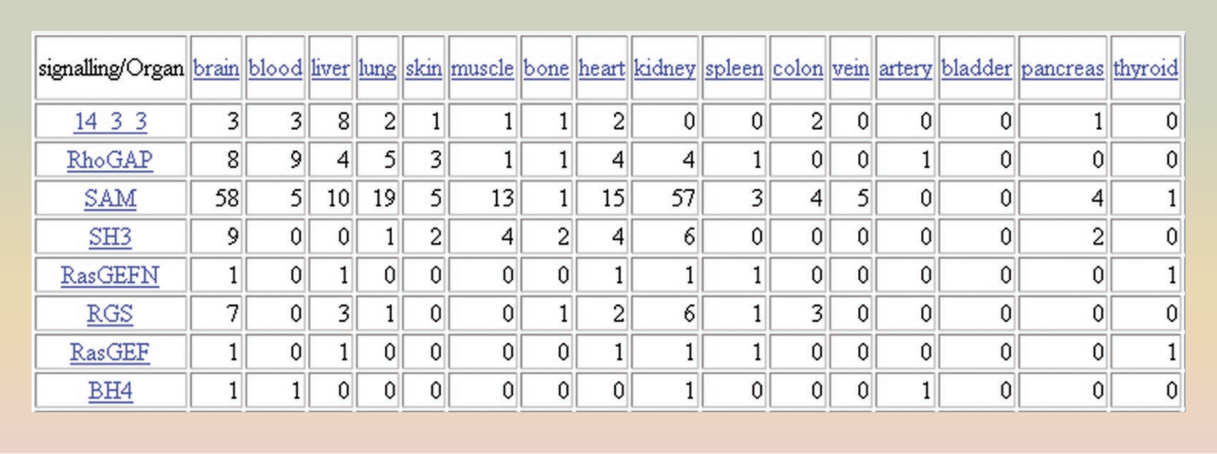


◉ cross category Mus musculus
○ vertical category Mus musculus
Cross Axle: ◉ By Frequecy
○ By Alphabet

Compound
Amino Acid
Organ
+Dry Lab Methods
+FUNCTION
+cellular component
+gene from Gene Ontology
+Root of LocusLink phenotype
+MeSH Minor (Tree)
+MeSH Major (Tree)
Minor MeSH
Major MeSH
+Protein By Species
Age Tags
Country
Data Completed
Publication Data
URL Full-Text
Last Revision Date
Organization Name
Prime Name of Substance
Personal Name as Subject
Publication Type
Subset
Secondary Source Identifier
Source
Summary For Patient In
Publication Status
Journal Title Abbreviation
Check Tags
Transliterated/Vernacular Title
Update In
Update Of
URL Summary
+PROPERTY
Minor Qualifier
Major Qualifier
+SACCHARIDE
+SMART Domain
Affiliation
Author
Data Created

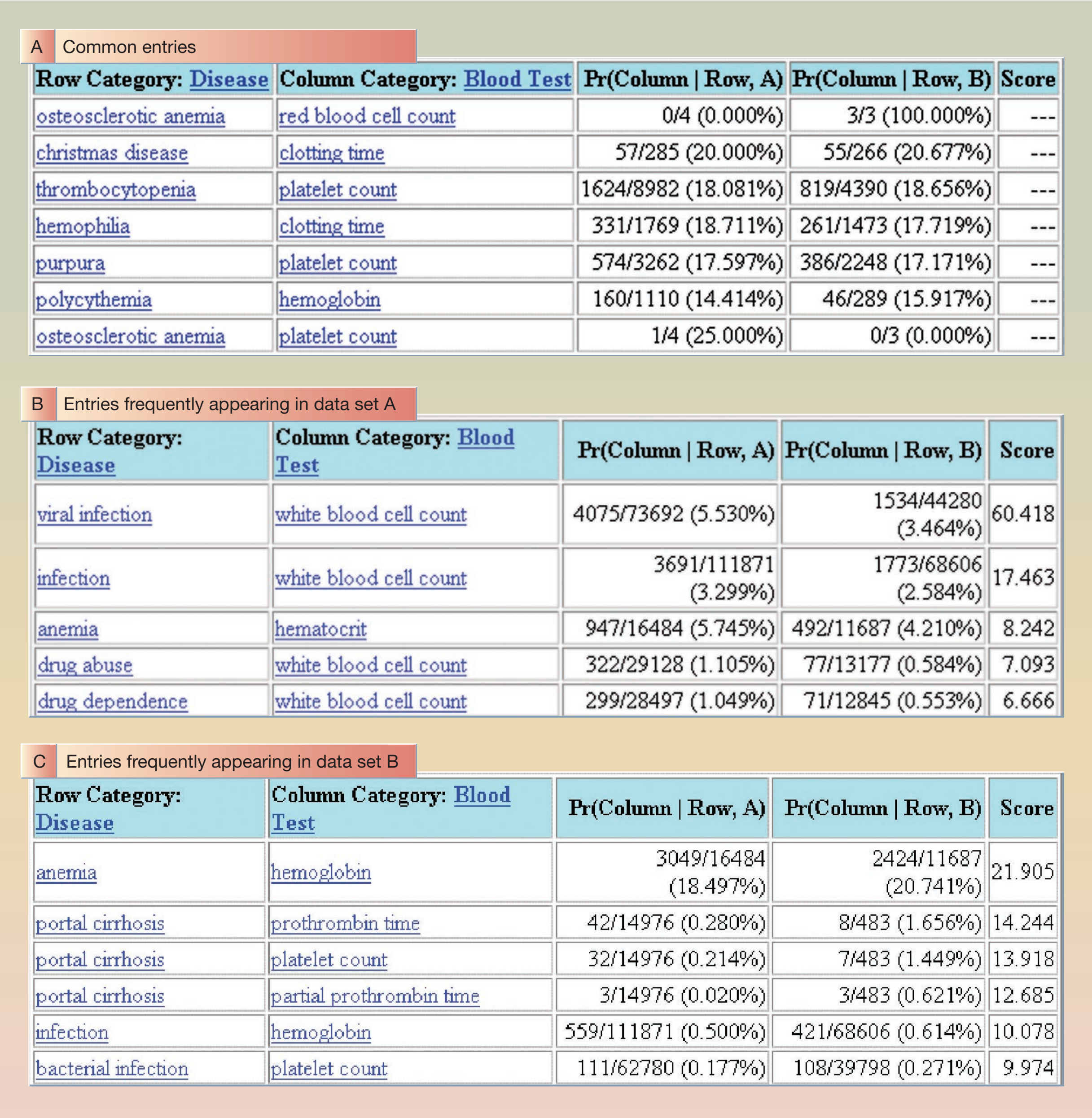
	protein	tumor suppressor	breast cancer	Bc12-associated X protein	myelocytomatosis oncogene	transcription factor
protein	632 (100.0%)	57 (9.01%)	42 (6.64%)	60 (9.49%)	24 (3.79%)	27 (4.27%)
tumor suppressor	57 (38.51%)	148 (100.0%)	4 (2.7%)	8 (5.4%)	6 (4.05%)	10 (6.75%)
breast cancer	42 (35.9%)	4 (3.41%)	117 (100.0%)	2 (1.7%)	8(6.83%)	3 (2.56%)
Bcl2-associated X-prot	60 (59.41%)	8 (7.92%)	2 (1.98%)	101 (100.0%)	4 (3.96%)	3 (2.97%)
myelocytomatosis onco	24 (38.1%)	6 (9.52%)	8 (12.7%)	4 (6.34%)	63 (100.0%)	2 (3.17%)
transcription factor	27 (45.0%)	10 (16.66%)	3 (5.0%)	3 (5.0%)	2 (3.33%)	60 (100.0%)
G elongation factor	28 (53.85%)	3 (5.76%)	1 (1.92%)	19 (36.54%)	1 (1.92%)	1 (1.92%)
proliferative cell nuclea	19 (37.25%)	1 (1.96%)	3 (5.88%)	2 (3.92%)	3 (5.88%)	1 (1.96%)
period	18 (37.5%)	2 (4.16%)	4 (8.33%)	2 (4.16%)	3 (6.25%)	0 (0.0%)
enhancer of rudimentar	17 (42.5%)	2 (5.0%)	23 (57.5%)	0 (0.0%)	1 (2.5%)	3 (7.5%)
epiregulin	17 (42.5%)	2 (5.0%)	23 (57.5%)	0 (0.0%)	1 (2.5%)	3 (7.5%)
progesterone receptor	13 (32.5%)	0 (0.0%)	20 (50.0%)	0 (0.0%)	1 (2.5%)	0 (0.0%)
Harvey rat sarcoma viru	12 (32.42%)	5 (13.51%)	0 (0.0%)	1 (2.7%)	9 (24.32%)	1 (2.7%)
tumor necrosis factor	11 (33.33%)	3 (9.09%)	0 (0.0%)	4 (12.12%)	1 (3.03%)	2 (6.06%)
epidermal growth factor	11 (33.33%)	2 (6.06%)	6 (18.18%)	0 (0.0%)	3 (9.09%)	0 (0.0%)
vascular endothelial gro	9 (29.03%)	0 (0.0%)	1 (3.22%)	1 (3.22%)	1 (3.22%)	1 (3.22%)
vitamin D receptor	10 (33.33%)	0 (0.0%)	2 (6.66%)	1 (3.33%)	1 (3.33%)	1 (3.33%)
estrogen receptor	8 (28.57%)	0 (0.0%)	16 (57.14%)	0 (0.0%)	1 (3.57%)	0 (0.0%)
protein-L-isoaspartate	7 (25.92%)	0 (0.0%)	2 (7.4%)	0 (0.0%)	3 (11.11%)	1 (3.7%)
MAS1 oncogene	9 (33.33%)	0 (0.0%)	11 (40.74%)	0 (0.0%)	1 (3.7%)	0 (0.0%)
protein kinase	7 (25.92%)	1 (3.7%)	0 (0.0%)	4 (14.81%)	1 (3.7%)	1 (3.7%)
cornichon	8 (30.76%)	1 (3.84%)	3 (11.54%)	0 (0.0%)	0 (0.0%)	1 (3.84%)

Figure 8 Result of similar entity search (query: 14_3_3)



signalling/Organ	brain	blood	liver	lung	skin	muscle	bone	heart	kidney	spleen	colon	vein	artery	bladder	pancreas	thyroid
14_3_3	3	3	8	2	1	1	1	2	0	0	2	0	0	0	1	0
RhoGAP	8	9	4	5	3	1	1	4	4	1	0	0	1	0	0	0
SAM	58	5	10	19	5	13	1	15	57	3	4	5	0	0	4	1
SH3	9	0	0	1	2	4	2	4	6	0	0	0	0	0	2	0
RasGEFN	1	0	1	0	0	0	0	1	1	1	0	0	0	0	0	1
RGS	7	0	3	1	0	0	1	2	6	1	3	0	0	0	0	0
RasGEF	1	0	1	0	0	0	0	1	1	1	0	0	0	0	0	1
BH4	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0

Figure 9 Result of subset analysis tool



A Common entries

Row Category: Disease	Column Category: Blood Test	Pr(Column | Row, A)	Pr(Column | Row, B)	Score
osteosclerotic anemia	red blood cell count	0/4 (0.000%)	3/3 (100.000%)	---
christmas disease	clotting time	57/285 (20.000%)	55/266 (20.677%)	---
thrombocytopenia	platelet count	1624/8982 (18.081%)	819/4390 (18.656%)	---
hemophilia	clotting time	331/1769 (18.711%)	261/1473 (17.719%)	---
purpura	platelet count	574/3262 (17.597%)	386/2248 (17.171%)	---
polycythemia	hemoglobin	160/1110 (14.414%)	46/289 (15.917%)	---
osteosclerotic anemia	platelet count	1/4 (25.000%)	0/3 (0.000%)	---

B Entries frequently appearing in data set A

Row Category: Disease	Column Category: Blood Test	Pr(Column | Row, A)	Pr(Column | Row, B)	Score
viral infection	white blood cell count	4075/73692 (5.530%)	1534/44280 (3.464%)	60.418
infection	white blood cell count	3691/111871 (3.299%)	1773/68606 (2.584%)	17.463
anemia	hematocrit	947/16484 (5.745%)	492/11687 (4.210%)	8.242
drug abuse	white blood cell count	322/29128 (1.105%)	77/13177 (0.584%)	7.093
drug dependence	white blood cell count	299/28497 (1.049%)	71/12845 (0.553%)	6.666

C Entries frequently appearing in data set B

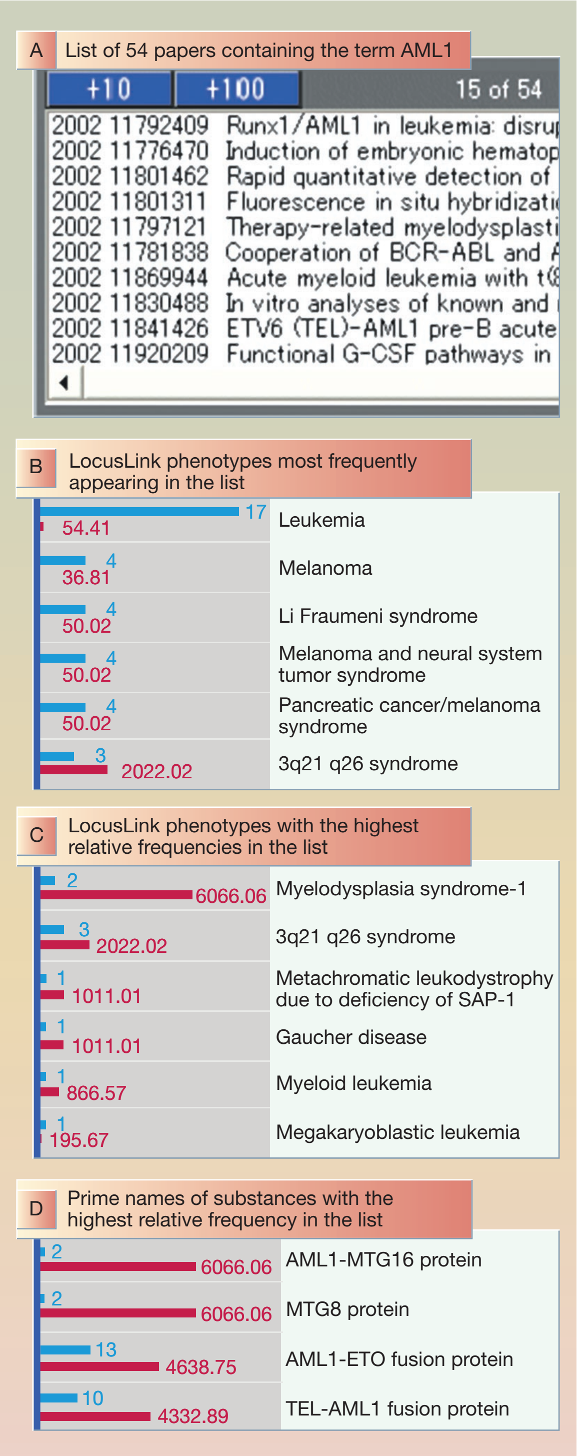
Row Category: Disease	Column Category: Blood Test	Pr(Column | Row, A)	Pr(Column | Row, B)	Score
anemia	hemoglobin	3049/16484 (18.497%)	2424/11687 (20.741%)	21.905
portal cirrhosis	prothrombin time	42/14976 (0.280%)	8/483 (1.656%)	14.244
portal cirrhosis	platelet count	32/14976 (0.214%)	7/483 (1.449%)	13.918
portal cirrhosis	partial prothrombin time	3/14976 (0.020%)	3/483 (0.621%)	12.685
infection	hemoglobin	559/111871 (0.500%)	421/68606 (0.614%)	10.078
bacterial infection	platelet count	111/62780 (0.177%)	108/39798 (0.271%)	9.974

$$\text{Ranking score} = -\log_{10} p$$

using the *p*-value calculated from the observed data.

Figure 9 shows the result of subset analysis by this function using disease-experiment two-dimensional-map data extracted by the MedTAKMI system. Documents are selected using the term "Adult" in the age category in data set A and "Child" in data set B. Figure 9A shows the common entries in both A and B. Figure 9B shows the entries which appear frequently in data set A, and Figure 9C shows those appearing frequently in data set B. In Parts 9B and 9C, entries are ranked by the statistical base score described earlier.

Figure 10 Results and analysis of keyword search for the term AML1



```
2002 11792409 Runx1/AML1 in leukemia: disru
2002 11776470 Induction of embryonic hematop
2002 11801462 Rapid quantitative detection of
2002 11801311 Fluorescence in situ hybridizati
2002 11797121 Therapy-related myelodysplasti
2002 11781838 Cooperation of BCR-ABL and A
2002 11869944 Acute myeloid leukemia with t(8
2002 11830488 In vitro analyses of known and
2002 11841426 ETV6 (TEL)-AML1 pre-B acute
2002 11920209 Functional G-CSF pathways in
```

B LocusLink phenotypes most frequently appearing in the list
17
54.41
Leukemia
4
36.81
Melanoma
4
50.02
Li Fraumeni syndrome
4
50.02
Melanoma and neural system tumor syndrome
4
50.02
Pancreatic cancer/melanoma syndrome
3
2022.02
3q21 q26 syndrome
C LocusLink phenotypes with the highest relative frequencies in the list
2
6066.06
Myelodysplasia syndrome-1
3
2022.02
3q21 q26 syndrome
1
1011.01
Metachromatic leukodystrophy due to deficiency of SAP-1
1
1011.01
Gaucher disease
1
866.57
Myeloid leukemia
1
195.67
Megakaryoblastic leukemia
D Prime names of substances with the highest relative frequency in the list
2
6066.06
AML1-MTG16 protein
2
6066.06
MTG8 protein
13
4638.75
AML1-ETO fusion protein
10
4332.89
TEL-AML1 fusion protein

Usage scenario

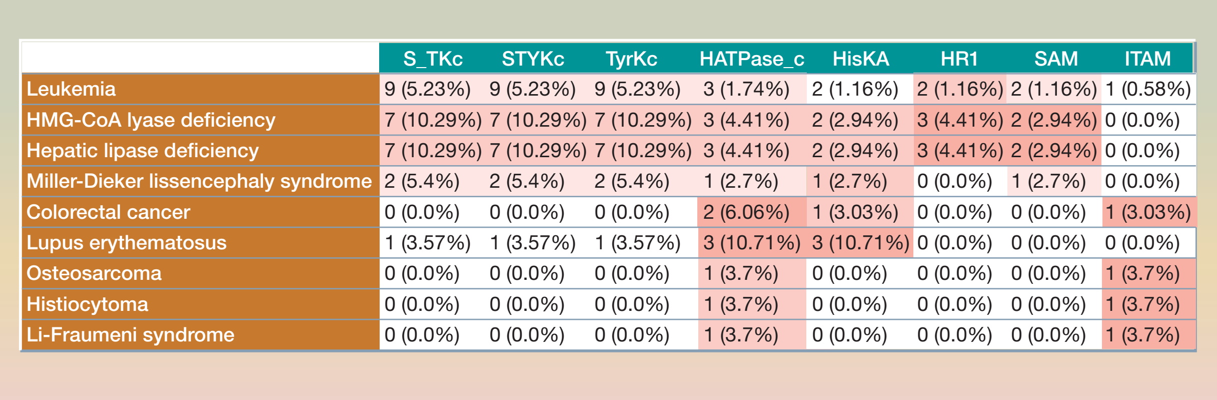
The MedTAKMI system can be used to aid in drug discovery and clinical information retrieval. For example, defects in the AML1 gene are thought to cause acute leukemia. Suppose we are interested in developing a treatment for leukemia and are looking for potential drug targets. We describe below the application of the MedTAKMI system to this problem. For this example a subset of approximately 330 000 of the latest abstracts from the 2003 MEDLINE distribution was used. A keyword search reveals that 54 papers in the sample set mention AML1. These are depicted in Figure 10. Each line shows the publication year, PubMed ID, and title of the paper.

A category view (Figure 10B) based on the NCBI LocusLink[44] phenotype information indicates that *leukemia* is indeed the term most frequently associated with AML1, appearing in 17 of the 54 papers. Furthermore, the relative frequency of the term *leukemia* indicates that it appears 54.41 times more often in this subset of 54 papers than it does in the entire sample database.

If we sort the entries in Figure 10B in decreasing order of relative frequency (Figure 10C), we see that myelodysplasia syndrome 1 (a protein that appears to be amplified in human prostate cancer)[45,46] and 3q21q26 syndrome (a syndrome associated with overexpression of the Evi-1 gene, an event in turn associated with myeloid leukemia) are the phenotypes that are strongly associated with AML1. Similarly, we can discover that various fusion genes and proteins such as AML1-MTG16, MTG8, AML1-ETO, and TEL-AML1 are also strongly associated with AML1 when we consider the category view for prime names of substances (Figure 10D).

Now, consider a search using the keyword *leukemia*, which selects 1051 papers from the entire sample database. We have already established that there appears to be a close association between AML1 and leukemia; if we can find another protein that is closely associated with leukemia, that protein might in turn serve as a potential target for drug design. A two-dimensional (2D) map analysis between LocusLink phenotypes and signaling proteins in this subcollection of 1051 papers shows very interesting results (Figure 11).

Figure 11 2D map analysis correlating LocusLink phenotypes (vertical axis) and signaling proteins (horizontal axis) in 1051 papers



	S_TKc	STYKc	TyrKc	HATPase_c	HisKA	HR1	SAM	ITAM
Leukemia	9 (5.23%)	9 (5.23%)	9 (5.23%)	3 (1.74%)	2 (1.16%)	2 (1.16%)	2 (1.16%)	1 (0.58%)
HMG-CoA lyase deficiency	7 (10.29%)	7 (10.29%)	7 (10.29%)	3 (4.41%)	2 (2.94%)	3 (4.41%)	2 (2.94%)	0 (0.0%)
Hepatic lipase deficiency	7 (10.29%)	7 (10.29%)	7 (10.29%)	3 (4.41%)	2 (2.94%)	3 (4.41%)	2 (2.94%)	0 (0.0%)
Miller-Dieker lissencephaly syndrome	2 (5.4%)	2 (5.4%)	2 (5.4%)	1 (2.7%)	1 (2.7%)	0 (0.0%)	1 (2.7%)	0 (0.0%)
Colorectal cancer	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (6.06%)	1 (3.03%)	0 (0.0%)	0 (0.0%)	1 (3.03%)
Lupus erythematosus	1 (3.57%)	1 (3.57%)	1 (3.57%)	3 (10.71%)	3 (10.71%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Osteosarcoma	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)
Histiocytoma	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)
Li-Fraumeni syndrome	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)

We find that some signaling proteins such as STYKc and TyrKc are commonly referred to in papers about leukemia, HMG-CoA lyase deficiency, hepatic lipase deficiency, and Miller-Dieker syndrome. Similarly, SAM is associated with the same phenotypes. (In fact, fusion of the SAM domain of the TEL oncogene to AML1 has been shown.[45,46]) In addition, HATPase_c appears associated with all the phenotypes. ITAM, on the other hand, might reveal a potential interesting relationship between leukemia and other phenotypes such as osteosarcoma. We can now click on a cell in the 2D map to access the corresponding papers if we are interested in pursuing a particular association. These results could then be explored as possible leads for drug development. Thus, the MedTAKMI system provides a contextual overview for identifying potential targets for drug design and facilitates the navigation of relevant papers.

Conclusions and future work

There are many useful applications for text mining in the life-science industry, particularly because of the vast amount of technical data and the myriad relationships contained therein that are waiting to be inferred, identified, and collated. There are also many textual databases other than MEDLINE to be explored. For example, laboratory notebooks, internal reports, and patent documents are all very important resources that contain valuable information about new gene, protein, and drug discoveries.

The next stages of the MedTAKMI system evolution include: (1) expanding the range of document sources to include these other databases, (2) developing a layer that allows integration of the MedTAKMI system with external and internal tools that are currently in daily use, and (3) improving the performance of the results system by developing a distributed version of MedTAKMI running on a grid architecture.[47,48] Because both the information-extraction process and the search/mining process handle several files independently, we believe that the file I/O and calculation costs can be shared across multiple processing units. If MedTAKMI can be extended to run on a grid architecture, the performance of both processes will be improved.

Compliance with the Unstructured Information Management Architecture (UIMA) proposed by a research community in IBM is also a major challenge.[49] The UIMA is a common framework for text analytics tools that are defined as text analysis engines (TAEs). We have already redefined our information extractors as TAEs. We are also working on developing text-mining middleware for life-science applications.[50] In this project, the MedTAKMI system runs as a part of UIMA-based middleware.

In summary, this paper has described the MedTAKMI text-mining system, a set of tools for knowledge discovery derived from millions of biomedical documents. The MedTAKMI system is able to parse 11 million MEDLINE citations with a syntactic shallow parser and extract relationships from these parsed entities with hierarchical category identifiers. It also provides a toolkit of interactive viewers to aid in the discovery of underlying knowledge in the literature of life science.

Acknowledgments

We would like to thank Celestar Lexico-Sciences, Inc. for collaborative work during the development and deployment of MedTAKMI.

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of the Gene Ontology Consortium, the United States National Library of Medicine, or Microsoft Corporation.

Cited references

1. S. Arlington, S. Barnett, S. Hughes, and J. Palo, *Pharma 2010: The Threshold of Innovation*, IBM Corporation, http://www-1.ibm.com/services/strategy/e_strategy/pharma_2010.html.
2. M. Hearst, "Untangling Text Data Mining," *Proceedings of the 37th Annual Meeting of the Association for Computer Linguistics (ACL'99)*, College Park, MD, ACL, East Stroudsburg, PA (1999), pp. 3–10.
3. Gene Ontology Consortium, http://www.geneontology.org.
4. D. Swanson, "Medical Literature as a Potential Source of New Knowledge," *Bulletin of the Medical Library Association* **78,** No. 1, 29–37 (1990).
5. V. Brusic and J. Zeleznikow, "Knowledge Discovery and Data Mining in Biological Databases," *Knowledge Engineering Review* **14,** No. 3, 257–277 (1999).
6. D. Swanson and N. Smalheiser, "An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery," *Artificial Intelligence* **91,** No. 2, 183–203 (1997).
7. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," *Proceedings of the 20th International Conference on Very Large Databases (VLDB)*, Santiago, Chile, ACM, New York (1994), pp. 487–499.
8. MEDLINE, http://www.nlm.nih.gov/databases/databases_medline.html.
9. P. Kankar, S. Adak, A. Sarkar, K. Murari, and G. Sharma, "MedMesh Summarizer: Text Mining for Gene Clusters," *Proceedings of the 2nd SIAM International Conference on Data Mining*, Arlington, VA (2002), pp. 548–565.
10. Medical Subject Headings (MeSH), http://www.nlm.nih.gov/mesh/meshhome.html.
11. L. Tanabe, U. Scherf, L. Smith, J. Lee, L. Hunter, and J. Weinstein, "MedMiner: An Internet Tool for Filtering and Organizing Gene Expression and Pharmacological Information," *BioTechniques* **27,** 1210–1217 (1999).
12. PubMed, http://www.ncbi.nlm.nih.gov/entrez.
13. T. Nasukawa and T. Nagano, "Text Analysis and Knowledge Mining System," *IBM Systems Journal* **40,** No. 4, 967–984 (2001).
14. K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, "Toward Information Extraction: Identifying Protein Names from Biological Papers," *Proceedings of the Pacific Symposium on Biocomputing '98 (PSB '98)*, Kapalua, HI, (1998), pp. 707–718.
15. M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman, "Using BLAST for Identifying Gene and Protein Names in Journal Articles," *Gene* **259,** No. 1–2, 245–252 (2000).
16. N. Collier, C. Nobota, and J. Tsujii, "Extracting the Names of Genes and Gene Products with a Hidden Markov Model," *Proceedings of the 18th International Conference on Computer Linguistics (COLING 2000)*, Saarbrücken, Germany, ACL, East Stroudsburg, PA (2000), pp. 201–207.
17. V. Hatzivassiloglou, P. Duboue, and A. Rzhetsky, "Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach," *Bioinformatics* **17,** Suppl. 1, S97–106 (2001).
18. K. Yamamoto, T. Kudo, A. Konagaya, and Y. Matsumoto, "Protein Name Tagging for Biomedical Annotation in Text," *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine (BioNLP)*, Sapporo, Japan, ACL, East Stroudsburg, PA (2003), pp. 65–72.
19. Y. Tsuruoka and J. Tsujii, "Boosting Precision and Recall of Dictionary-Based Protein Name Recognition," *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine (BioNLP)*, Sapporo, Japan, ACL, East Stroudsburg, PA (2003), pp. 41–48.
20. C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia, "Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions," *Proceedings of the AAAI Conference on Intelligent Systems in Microbiology (ISMB '99)*, Heidelberg, Germany, AAAI, Menlo Park, CA (1999), pp. 60–77.
21. T. Sekimizu, H. Park, and J. Tsujii, "Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in MEDLINE Abstracts," *Genome Informatics*, 62–71 (1998).
22. T. Rindflesch J. Rajah, and L. Hunter, "Extracting Molecular Binding Relationships from Biomedical Text," *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-NAACL 2000)*, Seattle, WA, ACL, East Stroudsburg, PA (2000), pp. 188–195.
23. A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, "Event Extraction from Biomedical Papers Using a Full Parser," *Proceedings of the Pacific Symposium on Biocomputing '01 (PSB'01)*, Kohala Coast, HI (2001), pp. 408–419.
24. C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles," *Bioinformatics* **17,** Suppl. 1, S74–82 (2001).
25. J. Park, H. Kim, and J. Kim, "Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorical Grammar," *Proceedings of the Pacific Symposium on Biocomputing '01 (PSB'01)*, Kohala Coast, HI (2001), pp. 396–407.
26. J. Hosaka, J. Koh, and A. Konagaya, "Effect of Utilizing Terminology on Extraction of Protein-Protein Interaction Information from Biomedical Literature," *Proceedings of the 10th Conference of the European Chapter of the Association for Computer Linguistics (EACL'03)*, Budapest, Hungary, ACL, East Stroudsburg, PA (2003), pp. 107–110.
27. S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, a Natural Language Processing Engine for MEDLINE Abstracts," *Bioinformatics* **19,** No. 13, 1699–1706 (2003).
28. J. Thomas, D. Milward, C. Ouzounis, and S. Pulman, "Automatic Extraction of Protein Interactions from Scientific Abstracts," *Proceedings of the Pacific Symposium on Biocomputing '00 (PSB'00)*, Honolulu, HI (2000), pp. 538–549.
29. K. Humphreys, G. Demetriou, and R. Gaizauskas, "Automatically Extracting Enzyme Interaction and Protein Structure Information from Biological Science Journal Articles," *Proceedings of the Symposium on Artificial Intelligence in Bioinformatics of the 2000 Convention of the Society for the Study*

of Artificial Intelligence and Simulation of Behavior (AISB'00), Birmingham, UK, AISB, Brighton, UK (2000).

30. K. Humphreys, G. Demetriou, and R. Gaizauskas, "Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures," *Proceedings of the Pacific Symposium on Biocomputing '00 (PSB'00)*, Honolulu, HI (2000), pp. 502–513.
31. G. Leroy and H. Chen, "Filling Preposition-Based Templates to Capture Information from Medical Abstracts," *Proceedings of the Pacific Symposium on Biocomputing '02 (PSB'02)*, Lihue, HI (2000), pp. 350–361.
32. M. Craven and J. Kumlien, "Constructing Biological Knowledge Bases by Extracting Information from Text Sources," *Proceedings of the AAAI Conference on Intelligent Systems in Microbiology (ISMB '99)*, Heidelberg, Germany, AAAI, Menlo Park, CA (1999), pp. 77–86.
33. E. Marcotte, I. Xenarios, and D. Eisenberg, "Mining Literature for Protein-Protein Interactions" *Bioinformatics* **17,** No. 4, 359–363 (2001).
34. See note 17.
35. B. Stapley and G. Benoit, "Biobibliometrics: Information Retrieval and Visualization from Co-occurrences of Gene Names in MEDLINE Abstracts," *Proceedings of the Pacific Symposium on Biocomputing '00 (PSB'00)*, Honolulu, HI (2000), pp. 529–540.
36. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/.
37. United States National Library of Medicine, http://www.nlm.nih.gov/.
38. MeSH Browser, http://www.nlm.nih.gov/mesh/MBrowser.html.
39. E. Charniak, *Statistical Language Learning*, MIT Press, Cambridge, MA (1993).
40. Unified Medical Language System, http://www.nlm.nih.gov/research/umls/umlsmain.html.
41. C. Pollard and I. Sag, *Head-Driven Phrase Structure Grammar*, University of Chicago Press and CSLI Publications, Chicago, IL (1994).
42. R. Berwick, Robert, "Computational Complexity and Lexical Functional Grammar," *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics*, Stanford, CA, ACL, East Stroudsburg, PA (1981), pp. 7–12.
43. H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control* **19,** 716–723 (1974).
44. LocusLink, http://www.ncbi.nih.gov/LocusLink/.
45. H. Tran, C. Kim, S. Faham, M. Siddall, and J. Bowie, "Native Interface of the SAM Domain Polymer of TEL," *BMC Structural Biology* **2,** No. 1, 5–11 (2002).
46. H. Sattler, R. Lensch, V. Rohde, E. Zimmer, E. Meese, H. Bonkhoff, M. Retz, T. Zwergel, A. Bex, M. Stoeckle, and B. Wullich, "Novel Amplification Unit at Chromosome 3q25-q27 in Human Prostate Cancer," *Prostate* **45,** No. 3, 207–15 (2000).
47. I. Foster, C. Kesselman, and S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *International Journal of Supercomputer Applications* **15,** No. 3, 200–222 (2001), http://www.globus.org/research/papers/anatomy.pdf.
48. *The Grid 2: Blueprint for a New Computing Infrastructure (2nd Edition)*, I. Foster and C. Kesselman, Editors, Morgan Kaufmann Publishing, Inc., San Francisco, CA (2003).
49. D. Ferrucci and A. Lally, "Building an Example Application with the Unstructured Information Management Architecture," *IBM Systems Journal* **43**, No. 3, 455-475 (2004, this issue).
50. R. Mack, S. Mukherjea, A. Soffer, N. Uramoto, E. Brown, A. Coden, J. Cooper, A. Inokuchi, B. Iyer, K. Kummamuru, Y. Maas, H. Matsuzawa, and L. Subramaniam, "Text Analytics for Life Science using the Unstructured Information Management Architecture," *IBM Systems Journal* **43**, No. 3, 490-515 (2004, this issue).

Accepted for publication February 6, 2004.

Naohiko Uramoto *IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato, Kanagawa, Japan (uramoto@jp.ibm.com).* Dr. Uramoto joined the IBM Tokyo Research Laboratory in 1990 after receiving a master's degree in computer science from Kyushu University. He received a Ph.D. in computer science from Kyushu University in 1990. He has been a visiting associate professor at the National Institute of Informatics since 2000. His research interests include natural language processing, text mining, XML and metadata, and information integration.

Hirofumi Matsuzawa *IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato, Kanagawa, Japan (matuzawa@jp.ibm.com).* Mr. Matsuzawa joined the IBM Tokyo Research Laboratory in 1993 after receiving a master's degree in software engineering from Waseda University. His experience includes 3D solid modeling, business intelligence, parallel OLAP, data mining, and text mining. His current focus is on data mining, text mining, and the application of grid architectures to life science.

Tohru Nagano *IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato, Kanagawa, Japan (tohru3@jp.ibm.com).* Mr. Nagano joined the IBM Tokyo Research Laboratory text mining project in 1998 after receiving a master's degree in computer science at the University of Tsukuba. Currently he is developing applications for text-mining systems. His research interests include natural language processing, machine learning, and statistical analysis.

Akiko Murakami *IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato, Kanagawa, Japan (akikom@jp.ibm.com).* Ms. Murakami joined the IBM Tokyo Research Laboratory in 1999 after receiving a master's degree in physics at Waseda University. Her current research interests include natural language processing and knowledge understanding in collaborative documents.

Hironori Takeuchi *IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato, Kanagawa, Japan (hironori@jp.ibm.com).* Mr. Takeuchi joined the IBM Tokyo Research Laboratory in 2000 after receiving a master's degree in mathematical engineering at the University of Tokyo. His research interests include information retrieval based on statistical analysis and patent knowledge management.

Koichi Takeda *IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato, Kanagawa, Japan (takedasu@jp.ibm.com).* Mr. Takeda joined the IBM Tokyo Research Laboratory in 1983 after receiving an M.E. degree in information science from Kyoto University. His research interests include machine translation, text analytics, and information retrieval.