MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling

BioTechniques 27:1210-1217 (December 1999)

L. Tanabe[1,2], U. Scherf[1], L.H. Smith[1], J.K. Lee[1], L. Hunter[1,2] and J.N. Weinstein[1]
[1]National Institutes of Health
Bethesda, MD, USA
[2]George Mason University
Fairfax, VA, USA

ABSTRACT

The trend toward high-throughput techniques in molecular biology and the explosion of online scientific data threaten to overwhelm the ability of researchers to take full advantage of available information. This problem is particularly severe in the rapidly expanding area of gene expression experiments, for example, those carried out with cDNA microarrays or oligonucleotide chips. We present an Internet-based hypertext program, MedMiner, which filters and organizes large amounts of textual and structured information returned from public search engines like GeneCards and PubMed. We demonstrate the value of the approach for the analysis of gene expression data, but MedMiner can also be extended to other areas involving molecular genetic or pharmacological information. More generally still, MedMiner can be used to organize the information returned from any arbitrary PubMed search.

INTRODUCTION

Given the current and projected growth of biomedical information on the Internet, we require Web-based tools that are capable of filtering the public databases and highlighting their relevant information in a well-organized and coherent manner. In particular, we found that we needed such tools to analyze gene-gene relationships observed in mRNA expression profiling experiments with cDNA microarrays and oligonucleotide chips. In analyzing relationships among thousands of genes, we found large numbers of apparent correlations—but were faced with the difficult task of determining which correlations represented interesting biological stories, which were simply epiphenomenal and which represented statistical coincidence. To address that question, we required fluent access to the best possible extrinsic information on the genes. However, the amount of information available on even a small subset of the thousands of genes proved too large to review using standard search engines. We therefore developed MedMiner, a computerized tool that filters the literature and presents the most relevant portions in a well-organized way that facilitates understanding. The result has been a considerable reduction in the time and effort required to survey the literature on genes and gene-gene relationships. The General Query option in MedMiner has proved similarly useful for any arbitrary PubMed search.

MedMiner was developed incrementally (with constant feedback from biologist-users to meet their needs). It incorporates several key computational components to achieve the twin goals of automated filtering and data organization. The first of these is Internet-based querying of multiple databases. Currently, our sources of information are PubMed and GeneCards (6), but the system is designed for easy integration of additional databases. The Materials and Methods section describes in detail the types of information extracted from these sources.

The second key component of MedMiner's procedure is text filtering. Text filtering systems translate user queries into relevance metrics that can be applied to large quantities of text automatically. Relevance metric research is an area of active investigation, but there are currently two widely used approaches. One applies combinations of keywords to identify relevant documents, paragraphs or sentences (4). The text filter might, for example, specify that an abstract is relevant if it contains a sentence with both the name of the gene and the word “inhibits”. A second approach uses word frequencies to determine relevance (4). A frequency-based filter might specify that an abstract is relevant if it contains words like “gene”, “inhibit” or “inhibition” significantly more frequently than does an average document. More sophisticated strategies for assessing relevance have also been proposed. Included are surface clue evaluation (1), shallow parsing (8), lexical and contextual analysis (5), semantic and discourse processing (2) and machine learning (11). However, the computational costs of applying any of these more sophisticated strategies to a large textual database is considerable, hence their use for Internet applications is problematical.

The third component of MedMiner is a carefully designed user interface. Because we are presenting large amounts of information, users must be able to navigate the material easily and modify their queries repeatedly to optimize results. The output is organized according to the relevance rule triggered, rather than being ordered arbitrarily or by date. This pattern

of organization makes browsing more logical and efficient.

While traditional information retrieval systems support user-formulated database queries for relevant documents, MedMiner searches documents for relevant facts specific to a predetermined domain. Our own studies have been done in the context of the National Cancer Institute's (NCI) drug discovery program, hence, we have been examining correlations between drug activity and gene expression (13). For that reason, MedMiner incorporates tools for literature exploration of gene-drug, as well as gene-gene, relationships. MedMiner (along with other analysis tools and genomic and pharmacological databases) is available at **http://discover.nci.nih.gov**, which can also be accessed through a link in the **Software Library** of the *BioTechniques* Web site (**www.biotechniques.com**). Although its development was motivated by our own needs with respect to gene expression profiling, it can easily be extended (without additional programming) to other pursuits (12) such as single nucleotide polymorphism (SNP) analysis, sequence analysis and proteomic profiling.

MATERIALS AND METHODS

Component Tools and Databases

MedMiner uses the Weizmann Institute's GeneCards database, mirrored at the NCI Web site (**http://nciarray.nci.nih.gov/cards/index.html**), and the National Library of Medicine's (NLM) PubMed database (**http://www.ncbi.nlm.nih.gov**) as its primary sources of data. The GeneCards database links information available in GDB, GenBank®, Swiss-Prot, OMIM, Medline and the Internet at large, and it is curated by human experts. A GeneCard entry is a highly condensed version of what is currently known about a particular gene. PubMed is a search tool for the NLM bibliographic database, which contains titles, citations, keywords and abstracts for most of the peer-reviewed scientific literature in biomedicine published from the 1960s to the present.

Processing Steps

Figure 1 shows schematically MedMiner's processing steps. The Common Gateway Interface (CGI) connects the client Web pages in hypertext markup language (HTML) format to Perl scripts running on a Web server. The programming language Perl was used because of its text processing capabilities and publicly available interface modules at **http://www.linpro.no/lwp/**.

The first step is for a user to specify genes of interest either by entering specific gene names (e.g., those located on a cDNA microarray or oligonucleotide chip) or by specifying a general concept (e.g., apoptosis) that can be used to find genes. The gene names or concepts are then sent as a query to the GeneCards database. MedMiner processes the result of the GeneCards query, highlighting the genes that are on the user-specified chip. These genes are annotated as being related to the query concept, and their names are augmented by including synonyms. Genes related to the query that are not on the chip are also displayed since they may be of interest for future chip design. For gene/gene queries, this process is

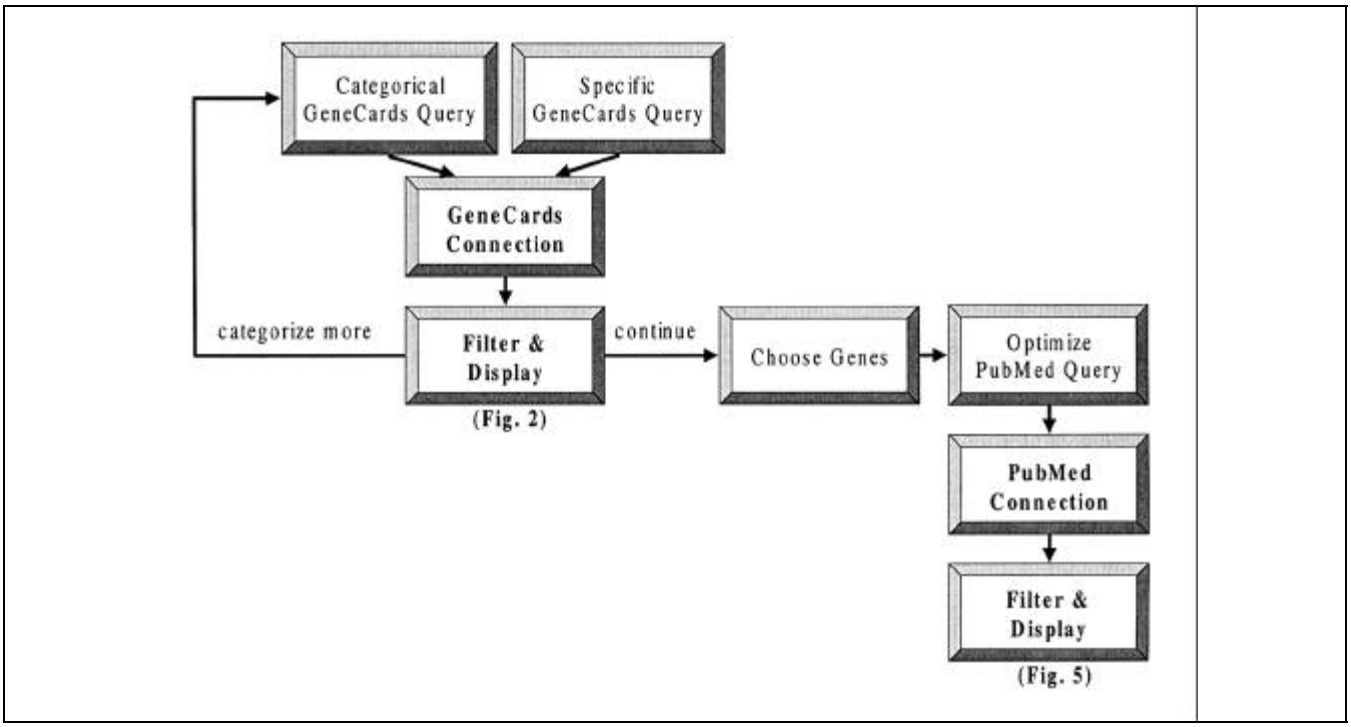

Figure 1. Flow chart of MedMiner component tools and databases. For simplicity, the figure illustrates the process for a gene query. Gene/gene and gene/drug queries are processed similarly. Plain lettering indicates a user input step; bold lettering indicates an automated step. For details, see the Materials and Methods section.

repeated for each of the two genes. For gene/drug queries, drug synonyms obtained from the NCI's Drug Information System database (**http://dtp.nci.nih.gov/**) are used.

The returned list can be filtered or modified either manually or automatically. Automatic refinement is currently restricted to matching this retrieval against the local gene database (e.g., the set of genes on an array). This step is particularly useful for gene expression microarray studies, in which we focus principally on genes that are actually on the array. The retrieved list (with any automatic filtering) is presented to the user, who can then refine the initial query or manually add or exclude other genes or concepts.

In the next step, the gene, drug and/or disorder names are formulated into a PubMed query. MedMiner creates a Boolean search with user-determined combinations of synonyms and retrieves the citations of matching articles. The user can reduce the size of the resulting retrieval by selecting a recent cutoff date or entering additional search terms. Even so, searching the entire biomedical literature for all articles relevant to a query can produce a long list of results. Hence, the number of abstracts found by PubMed is returned to the user before the filtering is performed.

The user then filters the results by applying the combinations-of-keywords method to the titles and abstracts. Our keyword combinations were devised to capture concepts relevant to the specific biomedical domain. For example, the relevance filters for genes involved in response to antitumor compounds look for terms like “inhibits”, “stimulates”, “resistance” and “sensitive.” Terms are truncated to approximate linguistic stemming (e.g., “inhibit” matches “inhibits” and “inhibition”). A sentence is considered relevant if it contains at least one gene synonym and at least one keyword. A citation is relevant if its title or abstract contains at least one relevant sentence.

Finally, citations that pass the relevance filter are grouped according to the particular relevance rule(s) triggered. For example, abstracts containing sentences about inhibition are grouped together. If a citation contains several relevant sentences, it will be represented in each of the corresponding groups. The results page shows the Boolean PubMed query, some summary statistics on the number of abstracts and sentences found, links to possible false positives (abstracts that MedMiner found irrelevant) and a set of hyperlinked tables. There are two options for navigating the tables—a heading distribution that groups similar keywords together on a graph (Figure 5A) and an alphabetical list that shows the frequency of each term separately—for example, “abnormal (1)” (Figure 5B). Each citation entry is annotated with the sentence that explains its relevance. The keywords are highlighted, and a link to the unfiltered abstract is provided.

Demonstration Problem

We use the human oncochip gene set provided by the NCI Microarray Facility (**http://nciarray.nci.nih.gov**) to show typical results. The initial query was the concept “DNA Repair”. Forty-five entries in GeneCards matched that search concept. MedMiner automatically compared these results with a local database containing the genes in the oncogene set and filtered the matches down to the 22 that are actually on the chip. The upper region of Figure 2 shows this result with genes on the chip highlighted in red, and each matching gene can be selected for further exploration with a checkbox. Genes relevant to the query but not on the chip are listed without highlighting. Figure 2 shows the set of all identification numbers for the relevant genes on the chip.

We use the well-known relationship between p53 and Mdm2 to illustrate the type of results obtained by filtering PubMed queries. Mdm2 regulates degradation of the p53 tumor suppressor protein (3) through an autoregulatory feedback loop. Enhanced p53 expression can be seen when the inhibitory effect of Mdm2 is decreased, for example, by c-Abl protein-tyrosine kinase (10).

RESULTS

To compare PubMed and MedMiner, we checked both the quality of the retrievals and the utility of the results. For the quality of retrieval, we compared MedMiner results with the results from hand-crafted PubMed queries. First, we compared the number of relevant abstracts returned from the PubMed query “p53 AND mdm2 AND inhibit*” (the asterisk indicates term explosion by the search engine) with the

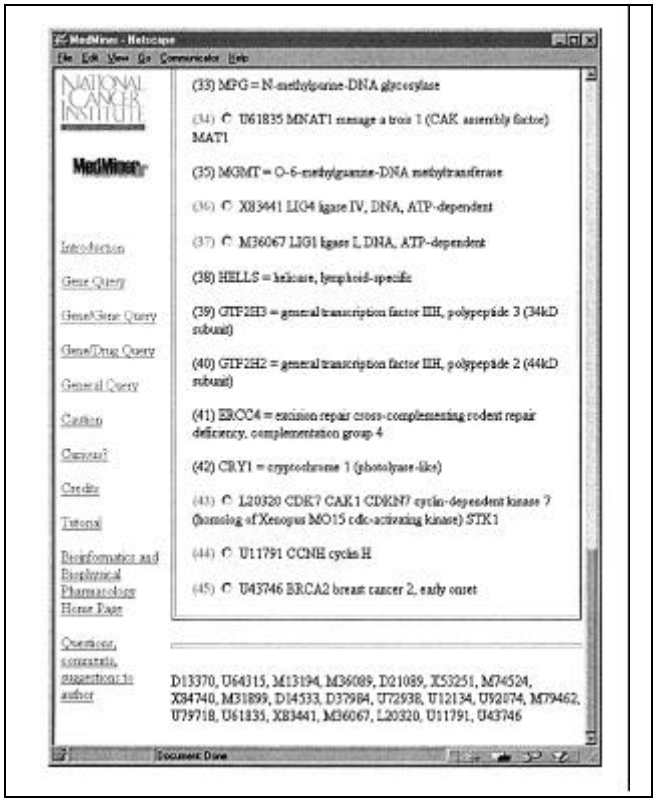

Figure 2. Finding genes related to DNA repair on an array. The query, “DNA repair”, returns relevant GeneCards, which are coupled by a gene identifier to the genes on the array. These genes, highlighted in red, are accompanied by their synonyms from a precompiled list. Genes not on the array are listed by name only. The comma-delimited list of identifiers at the bottom of the page is used to access local gene expression data. This Web page will evolve as new capabilities are added.

number of relevant abstracts returned by the MedMiner query "p53, mdm2", both with a 60-day cutoff. Because the relationship between p53 and mdm2 is well understood, we focused on the "inhibition" relationship. PubMed retrieved 12 citations for the query "p53 AND mdm2 AND inhibit*", and MedMiner returned 11 citations that it determined were relevant to p53, mdm2 and inhibition.

This one-abstract difference is instructive. The extra sentence containing "inhibit" was: "The mammalian ARF-INK4a locus uniquely encodes two cell cycle inhibitors by using separate promoters and alternative reading frames" (14). PubMed included the abstract even though it did not indicate an inhibitory relationship between p53 and MDM2, whereas MedMiner appropriately excluded it from the "inhibit" group of citations. However, the same abstract also included the sentence: "ARF forms nuclear bodies with MDM2 and p53 and blocks p53 and MDM2 nuclear export", so MedMiner correctly included the citation under the keyword "block".

We note that the most common way to scan a large number of PubMed results is by title. Of the 12 citations returned by the PubMed search on p53, Mdm2 and inhibition, only two contained the word "inhibit" in the title. To determine why the remaining 10 citations were retrieved, a user would have to read the 10 abstracts. In contrast, the MedMiner table for "inhibition" showed the sentence from each abstract that asserted the inhibition relationship, highlighting the keywords and the gene names (see Figure 3).

Considering all of the MedMiner relevance filters together (not just inhibition), the next question we addressed was whether reasonable manually constructed PubMed queries could be made as specific as MedMiner queries. The MedMiner query, "p53, Mdm2", generated 31 relevant abstracts and organized them by topic. To generate the same specificity in a PubMed query, the shortest equivalent query we could devise was "p53 AND Mdm2 AND (inhibit* OR mutat* OR block* OR report* OR induc* OR tumor OR reduc* OR result* OR suggest* OR regulat* OR depend* OR mediat* OR indicat* OR essential OR promot* OR express* OR low OR accumulat* OR increas* OR activat* OR effect OR involv* OR activit* OR requir* OR control OR level OR respons* OR stimulat* OR role OR presence OR interact OR propose OR deficien* OR decreas* OR apopto* OR sensitiv* OR abnormal OR tumour OR high OR absence OR cleav* OR negative OR positive OR proliferat* OR correlat* OR malignant OR enhanc* OR critical OR rise OR evidenc* OR shown OR cataly* OR influenc* OR signif* OR conclud* OR identif* OR agent)". The one abstract that only PubMed returned as relevant contained the sentence: "To evaluate which molecular biological factors are related to patients' prognosis and recurrence, we checked p53, p16, p21/Waf1, cyclin D1, Ki-67, epidermal growth factor receptor (EGFR), vascular endothelial growth factor (VEGF), Mdm2, Bcl2, E-cadherin and MRP1/CD9 by means of immunohistochemical analysis in 116 cases of oesophageal cancer (R0)" (9). The abstract was returned by PubMed because it contained the query terms "role", "reduc*" and "indicat*". However, these relationships did not involve p53 or Mdm2, so MedMiner's omission of this citation was appropriate.

In terms of running speed, there is a moderate penalty for using MedMiner. PubMed retrieval for this query over the Internet took approximately 30 s. MedMiner's total processing time, including posting of the initial query, choosing of

	PubMed only	MedMiner
Query	p53 AND mdm2 AND inhibit*	p53,mdm2
Cutoff date	past 60 days	past 60 days
Number of abstracts returned as relevant	12	11
Initial results format	Abstract titles	Abstract sentences
Number of initial results	12	17
Number of initial results with "inhibit" relationship	2/12	17/17
Number of sentences returned as relevant	104	17

Figure 3. Comparison of results from PubMed and MedMiner for a query on inhibitory p53 and MDM2 relationships. See the Results section for a detailed description. The 60-day cutoff is relative to the date the queries were made, hence the results will be different if the same comparison is done at a later date.

	PubMed only	MedMiner
Query	P53 AND mdm2 AND mutat*...OR...agent	P53, mdm2
Cutoff date	Past 60 days	Past 60 days
Initial results format	Abstract titles	Abstract sentences
Number of abstracts returned as relevant	32	31
Number of initial results	32	180
Processing time	30 seconds	62 seconds
Number of sentences returned as relevant	395	180
Highlighted query terms	N	Y
Organized by keyword	N	Y

Figure 4. Comparison of results from PubMed and MedMiner for a query on multiple p53 and MDM2 relationships. See the Results section for a detailed description. The 60-day cutoff is relative to the date the queries were made, hence the results will be different if the same comparison is done at a later date.

synonyms, processing of text and completion of Web presentation was slightly more than 60 s.

Although the number of abstracts retrieved by such a large PubMed query is comparable to the number found by MedMiner, there is a major difference in their organization. It is generally impossible to determine by the title of the citation which relationship caused it to be retrieved. For the 32 abstracts retrieved, there were 395 sentences in the abstracts, each of which would need to be read to determine the relevance of the abstracts. In contrast, MedMiner found 180 sentences that had at least one of the 51 keywords along with a gene identifier. In addition to highlighting relevant sentences (rather than titles), MedMiner also sorted the citations into alphabetized keyword categories (see Figure 4).

DISCUSSION

MedMiner offers a potentially significant new aid for coping with the torrent of molecular biology data confronting today's scientist. By filtering and organizing material retrieved from high-quality Internet sites, this program makes complex database searches much easier to execute and digest. Figure 2 illustrates that MedMiner successfully integrates public databases with local ones by using the local database to filter the much larger public ones. This link is important to save time and keep the local database synchronized with the constantly updated public ones. Additional databases could be merged into the system, integrating a wider variety of filters with a consistent user interface. The relevance keyword list is simple to create and update. Hence, MedMiner could be applied to additional biomedical domains such as comparative genomics, cytogenetics, proteomics and SNP analysis simply by compiling a list of relevant keywords specific to the domain. No additional programming would be required. The computational resources required for this kind of text processing are moderate, so the tool can be provided to a larger user community through the World Wide Web. For example, MedMiner has been incorporated into the Mouse Oncochip Design tool, a Web site designed to obtain mouse DNA sequences expressed

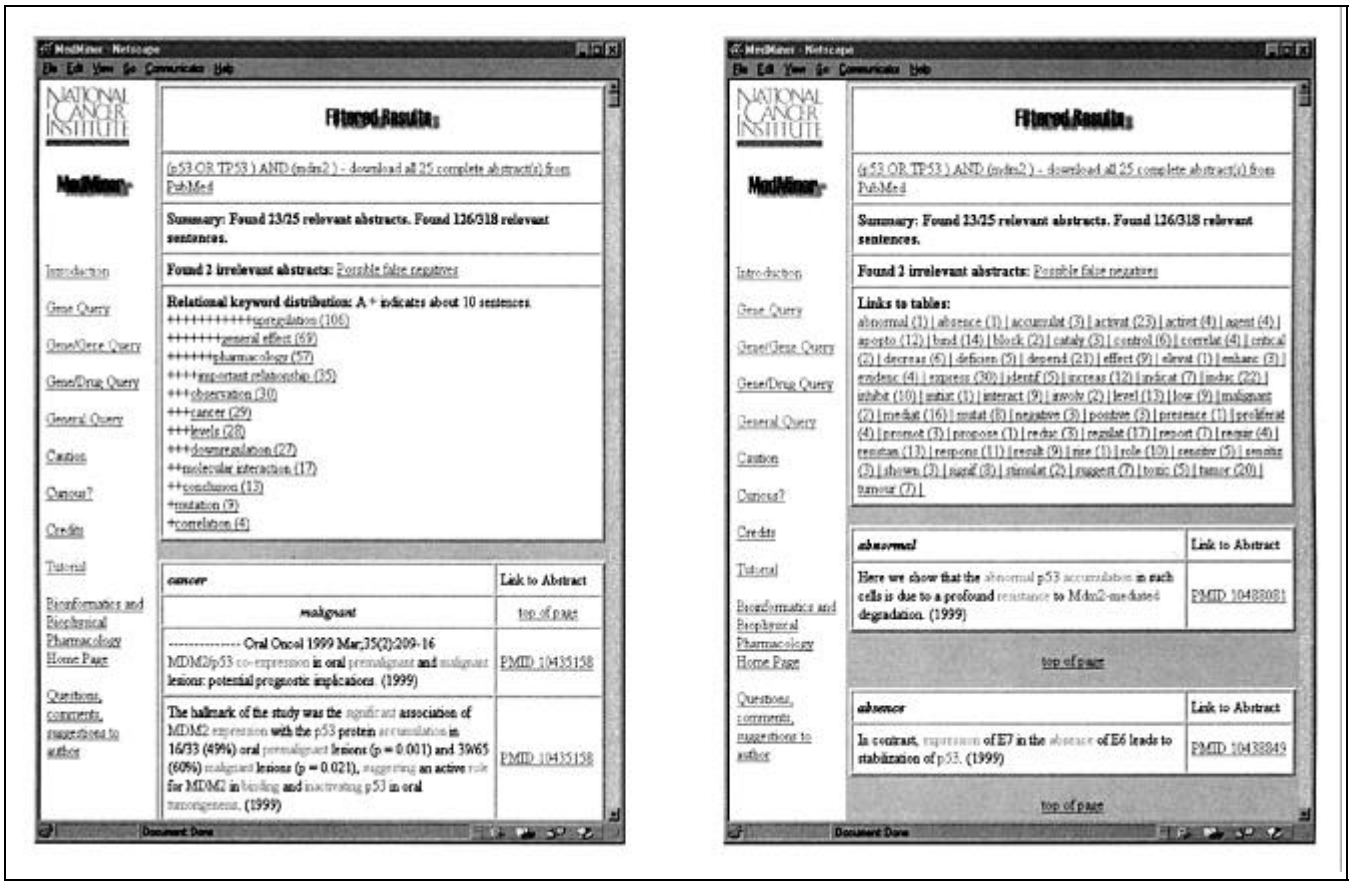

Figure 5. MedMiner results for p53 and Mdm2. The link "(P53) and (MDM2)-download32 complete abstracts from PubMed", at the top of the page will bring the user to unfiltered PubMed abstracts based on the optimized query. This link is followed by a short summary indicating how many sentences and abstracts were determined to be relevant. (A) Navigation using heading distribution. A frequency distribution illustrates the number of relevant sentences under each heading along with links to the corresponding tables. For example, the "cancer(29)" link brings the user to the table of 29 sentences containing any of the cancer-related keywords "malignan", "tumour", "tumor" and/or "neoplas". The full abstract can be viewed by following the appropriate link. Other heading relationships are navigated similarly. (B) Navigation using alphabetical keywords list. A list of alphabetical keyword hyperlinks shows the number of relevant sentences found for each keyword. For example, the "abnormal(1)" link brings the user to the table of one sentence containing the keyword "abnormal". The full abstract can be viewed by following the appropriate link. Other keyword relationships are navigated similarly. These Web pages will evolve as new capabilities are added.

in specific tissues for use in transcript microarray design (**http://nciarray.nci.nih.gov/cgi-bin/me/mouse_design.cgi**).

However, there are several limitations to this approach. The most important arises from the use of a keyword list to identify more general concepts. The program will miss relevant concepts if they are not represented in the keywords. For example, inhibition relationships are paraphrased in multiple ways in abstracts, but MedMiner will find only those relationships that explicitly contain some form of the word "inhibit". If the inhibitory relationship is expressed in terms of a down-regulation, it will be missed unless some form of the term "down-regulation" is also on the keywords list. Another problem is that our lexical analysis is quite simple. We use simple truncation to recognize linguistic variations, but not all such variations are amenable to this approach. Perhaps the most significant shortcoming of the keyword approach is that it will not pick up relationships that are specified across multiple sentences. For gene/gene or gene/drug queries, the system considers a sentence relevant if it contains at least one gene or drug synonym plus a keyword.

The most important question about this system is also the hardest to answer: How much time does it save? The Results section states that the processing time for a typical MedMiner query is moderately longer than that for a comparable PubMed query. However, this time investment is more than offset by the overall time saved in reading and interpreting the structured set of sentences, rather than the full abstracts. One rough estimate can be made by observing investigators analyzing gene expression profiles in our laboratory. Although PubMed and MedMiner are equally available (just by selecting Web pages), the investigators have largely opted to use MedMiner. An entirely subjective estimate by one of the investigators was that preparing and digesting the results of a complex search was ten times faster with MedMiner. The difference resulted in part from the easy identification of synonyms and in part from filtering using the local gene database, but principally from the more transparent organization of the retrieved citations.

We have found that MedMiner provides an efficient, empirically based, domain-dependent, first-pass filter of the vast amount of textual data currently available for gene expression profiling. To complement our combination-of-keywords approach with more sophisticated semantic and lexical analyses, we are collaborating with the Semantic Knowledge Representation group at NLM (7).

ACKNOWLEDGMENTS

We thank Mary Edgerton and Ronald C. Taylor for helpful comments and suggestions. This work was partially supported by an award from the National Cancer Institute Breast Cancer Think Tank.

REFERENCES

1. **Fukuda, K., T. Tsunoda, A. Tamura and T. Takagi.** 1998. Toward information extraction: protein names from biological papers. Proc. Pacific Symposium on Biocomputing *3*:705-716.
2. **Hishiki, T., N. Collier, C. Nobata, T. Okazaki-Ohta, N. Ogata, T. Sekimizu, R. Steiner, H.S. Park and J. Tsujii.** 1998. Developing NLP tools for genome informatics: An information extraction perspective. Proc. Ninth Workshop on Genome Informatics: 81-90.
3. **Kubbutat, M.H., R.L. Ludwig, A.J. Levine and K.H. Vousden.** 1999. Analysis of the degradation function of Mdm2. Cell Growth Differ. *10*:87-92.
4. **Oard, D.W.** 1997. The state of the art in text filtering. UMUAI: 141-178.
5. **Proux, D., F. Rechenmann, L. Julliard, V. Pillet and B. Jacq.** 1998. Detecting gene symbols and names in biological texts: A first step toward pertinent information. Proc. Ninth Workshop on Genome Informatics: 72-80.
6. **Rebhan, M., V. Chalifa-Caspi, J. Prilusky and D. Lancet.** 1997. GeneCards: encyclopedia for genes, proteins and diseases, Weizmann Institute of Science, Bioinformatics Unit and Genome Center.
7. **Rindflesch, T.C., L. Tanabe, J.N. Weinstein and L. Hunter.** 2000. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Proc. Pacific Symposium on Biocomputing *5*:514-525.
8. **Sekimizu, T., H.S. Park and J. Tsujii.** 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. Proc. Ninth Workshop on Genome Informatics: 62-71.
9. **Shimada, Y., M. Imamura, G. Watanabe, S. Uchida, H. Harada, T. Makino and M. Kano.** 1999. Prognostic factors of oesophageal squamous cell carcinoma from the perspective of molecular biology. Br. J. Cancer *80*:1281-1288.
10. **Sionov, R.V., E. Moallem, M. Berger, A. Kazaz, O. Gerlitz, Y. Ben-Neriah, M. Oren and Y. Haupt.** 1999. c-Abl neutralizes the inhibitory effect of Mdm2 on p53. J. Biol. Chem. *274*:8371-8374.
11. **Usuzaka, S., K.L. Sim, M. Tanaka, H. Matsuno and S. Miyano.** 1998. A machine learning approach to reducing the work of experts in article selection from database: A case study for regulatory relations of S. cerevisiae genes in MEDLINE. Proc. Ninth Workshop on Genome Informatics: 91-101.
12. **Weinstein, J.N.** 1998. Fishing Expeditions. Science *282*:627.
13. **Weinstein, J.N., T.G. Myers, P.M. O'Connor, S.H. Friend, A.J. Fornace, K.W. Kohn, T. Fojo, S.E. Bates et al.** 1997. An information-intensive approach to the molecular pharmacology of cancer. Science *275*:343-349.
14. **Zhang, Y. and Y. Xiong.** 1999. Mutations in human ARF exon 2 disrupt its nucleolar localization and impair its ability to block nuclear export of MDM2 and p53. Mol. Cell. *3*:579-591.

Received 20 July 1999; accepted 7 October 1999.

Address correspondence to:

Lorraine Tanabe
37 Convent Dr.
Bldg. 37, Rm. 5B12
Bethesda, MD 20892, USA
Internet: ltanaben@molstat.nci.nih.gov