Text Influenced Molecular Indexing (TIMI): A Literature Database Mining Approach that Handles Text and Chemistry

Suresh B. Singh,* Richard D. Hull,† and Eugene M. Fluder

Department of Molecular Systems, Merck Research Laboratories, 126 East Lincoln Avenue, RY50SW-100, Rahway, New Jersey 07065-0900

Received August 20, 2002

We present an application of a novel methodology called Text Influenced Molecular Indexing (TIMI) to mine the information in the scientific literature. TIMI is an extension of two existing methodologies: (1) Latent Semantic Structure Indexing (LaSSI), a method for calculating chemical similarity using two-dimensional topological descriptors, and (2) Latent Semantic Indexing (LSI), a method for generating correlations between textual terms. The singular value decomposition (SVD) of a feature/object matrix is the fundamental mathematical operation underlying LSI, LaSSI, and TIMI and is used in the identification of associations between textual and chemical descriptors. We present the results of our studies with a database containing 11 571 PubMed/MEDLINE abstracts which show the advantages of merging textual and chemical descriptors over using either text or chemistry alone. Our work demonstrates that searching text-only databases limits retrieved documents to those that explicitly mention compounds by name in the text. Similarly, searching chemistry-only databases can only retrieve those documents that have chemical structures in them. TIMI, however, enables search and retrieval of documents with textual, chemical, and/or text- and chemistry-based queries. Thus, the TIMI system offers a powerful new approach to uncovering the contextual scientific knowledge sought by the medical research community.

INTRODUCTION

A fundamental component of medical research, or any research effort for that matter, is understanding the prior art of the field. Consequently, great effort is expended in finding and integrating relevant literature references and concepts into a knowledge base to support current and ongoing research efforts. This is especially challenging for medicinal chemists, whose domain of interest involves both chemistry (chemical structures) and biology (textual descriptions of biological systems and their inter-relationships). Researchers are often challenged by the need to search the literature using both chemical structures and words. The majority of literature search facilities, however, support only keyword searches. Literature references retrieved by these search facilities are useful but are critically dependent on the keywords used. Other search facilities provide substructure searching, but during the early stages of a drug discovery project it is unlikely that a pharmacophore or the mechanism of action is known. Neither of these types of search facilities exploits the contextual environment in which the keywords exist in the retrieved articles. Moreover, medicinal chemists would often like to query literature sources with chemical structures in place of or in addition to keywords. A novel methodology, called Text Influenced Molecular Indexing (TIMI), was developed with these issues in mind. TIMI can handle a variety of query terms, including chemical structure(s) to mine the scientific literature.

* Corresponding author phone: (732)594-4954; fax: (732)594-4224; e-mail: suresh_singh@merck.com, singhsu@yahoo.com. Corresponding author address: RY50SW 100, P.O. Box 2000, Rahway, NJ 07065.

† Current address: Hull Consulting, Inc., 2646 Windsorgate Lane, Orlando, FL 32828.

The early stages of a drug discovery project are dedicated to finding "lead" compounds, i.e., compounds that can lead the project to an eventual drug. Lead compounds are often identified by various experimental and *in silico* screening processes. *In silico* approaches to chemical database screening have become a foundation of the drug industry because of the speed and reliability of these kinds of searches and also due to the dramatic increase in the size of most commercial and proprietary compound collections over the past decade.[1]

Many strategies for representing molecules in the collection and computing similarity between them have been devised;[2,3] however, none of these schemes to date take advantage of a crucial component of similarity: the textual context that exists in the medicinal chemistry literature describing chemical compounds. Our hypothesis is that there are meaningful relationships between compounds that are not just encoded in their structures but can be found in the textual descriptions surrounding them in the medical literature. Identifying these relationships may provide new insights into the behavior of molecules *in vivo*.

We developed TIMI as an extension to our recently published novel methodology for calculating chemical similarity using two-dimensional topological descriptors called Latent Semantic Structure Indexing, or LaSSI.[4−6] Both TIMI and LaSSI were inspired by the work of Deerwester et al.[7] on Latent Semantic Indexing (LSI). The singular value decomposition (SVD) of a feature/object matrix is the cornerstone of all three techniques and is used in the identification of associations between text and chemistry.

The remainder of this article discusses the origins of TIMI and how the methodology works. An example involving a modest number of MEDLINE abstracts is then presented and discussed.

10.1021/ci025587a CCC: $25.00
Published on Web 04/19/2003

THEORY: LSI AND LaSSI

The idea of TIMI arose from an initial investigation of LSI and later our work on LaSSI.

The mathematical underpinnings of LaSSI, presented elsewhere,[4] were inspired by Latent Semantic Indexing (LSI), an information retrieval technique originally described by Deerwester et al.[7] LSI represents a collection of text documents as a term-document matrix. The ultimate purpose is retrieving documents from a corpus given a user's query. LaSSI, on the other hand, uses a chemical descriptor-molecule matrix to calculate chemical similarities. Hence, the nature of the input matrices for LaSSI and LSI are very different. The mathematical treatment of these matrices, however, is the same. Later, we will see that the calculation of object similarities made by LSI and LaSSI is related but different.

In LASSI, a collection of molecules in a chemical database is initially represented as a set of vectors, where each vector $V_i = (\boldsymbol{d_{i1}}, \boldsymbol{d_{i2}}, \boldsymbol{...}, \boldsymbol{d_{in}})^{\mathrm{T}}$ consists of the nonnegative frequency of occurrence of each descriptor d_j in molecule i and where n is the total number of uniquely occurring descriptors in the entire set. A chemical descriptor-molecule matrix, X, therefore, is a set of two or more such vectors, i.e., $X = \{V_1, ..., V_m\}$, m $\geq$ 2, or

$$X = \begin{array}{c} \text{molecules} \\ \begin{bmatrix} d_{11} & d_{21} & \cdots & d_{m1} \\ d_{12} & d_{22} & \cdots & d_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ d_{1n} & d_{2n} & \cdots & d_{mn} \end{bmatrix} \end{array} \text{descriptors}$$

LaSSI employs the SVD of X to produce a reduced dimensional representation of the original matrix. The SVD technique is well known in the linear algebra literature[8] and has been used in many scientific and engineering applications including signal and spectral analysis.[9−11] Previously we showed a novel application of SVD to the problem of calculating chemical similarity.[4,5]

Let the SVD of X in $\mathrm{R}^{m\times n}$ be defined as $X = P\Sigma Q^T$ where P is an $n\times r$ matrix, called the left singular matrix (r is the rank of X), and its columns are the eigenvectors of XX^T corresponding to nonzero eigenvalues. Q is an $m\times r$ matrix, called the right singular matrix, whose columns are the eigenvectors of X^TX corresponding to nonzero eigenvalues. Σ is an $r\times r$ diagonal matrix = diag(σ_1, σ_2, ..., σ_r) whose nonzero elements, called singular values, are the square roots of the eigenvalues and have the property such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$.

$$\begin{array}{c} X \\ \begin{bmatrix} d_{11} & d_{21} & \cdots & d_{m1} \\ d_{12} & d_{22} & \cdots & d_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ d_{1n} & d_{2n} & \cdots & d_{mn} \end{bmatrix} \end{array} \xrightarrow{\text{SVD}} \begin{array}{c} P \\ \begin{bmatrix} P_{11} & \cdots & P_{r1} \\ P_{12} & \cdots & P_{r2} \\ \vdots & \vdots & \vdots \\ P_{1n} & \cdots & P_{rn} \end{bmatrix} \end{array} \bullet$$

$$\begin{array}{c} \Sigma \\ \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \end{array} \bullet \begin{array}{c} Q^{\mathbf{T}} \\ \begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ Q_{r1} & Q_{r2} & \cdots & Q_{rm} \end{bmatrix} \end{array}$$

The kth rank approximation of X, X_k, for $k < r$, σ_{k+1} ... σ_r set to 0 can be efficiently computed using variants of the Lanczos algorithm.[12] X_k is the matrix of rank k which is the closest to X in the least squares sense, is called a partial SVD of X, and is defined as $X_k = P_k\Sigma_kQ^T_k$.

Deerwester et al.[7] showed that given the partial SVD of X, it is possible to compute similarities between language terms, between documents, and between a term and a document. Furthermore, they could compute the similarity of *ad hoc* queries (column vectors which do not exist in X) to both the terms and the documents in the database. We now apply these computations to the chemical domain. The similarity of two descriptors, D_i and D_j, is calculated by computing the dot product between the ith and jth rows of the matrix $P_k\Sigma_k$. The similarity of two molecules, represented by vectors V_i and V_j, can be calculated by computing the dot product between the ith and jth rows of the matrix $Q_k\Sigma_k$. The similarity of a descriptor, D_i, to a molecule, V_j, can be calculated by computing the dot product between the ith row of the matrix $P_k\Sigma_k$ and the jth row of the matrix $Q_k\Sigma_k$. Finally, the similarity of an ad hoc query to the descriptors and molecules in the database can be calculated by first projecting the query into the k-dimensional space of the partial SVD and then treating the projection as a molecule for between and within comparisons. The projection of a query vector, v, is defined as $y = v^TP_k\Sigma^{-1}{}_k$.

LaSSI does not use the singular values to scale the singular vectors, however, as is the case for LSI. Instead, the identity matrix I is used in place of Σ_k when calculating similarities. The chemical similarity calculations were more meaningful when we did not scale the singular vectors with the singular values.[5] Therefore, the calculation of LaSSI similarity between two descriptors, two molecules, and a molecule and a descriptor is shown below.

LaSSI similarity between two descriptors

$$D_i \text{ and } D_j = \sum_{x=1}^{k} \frac{Pix}{|Pi|} \cdot \frac{Pjx}{|Pj|}$$

LaSSI similarity between two molecules

$$V_i \text{ and } V_j = \sum_{x=1}^{k} \frac{Qix}{|Qi|} \cdot \frac{Qjx}{|Qj|}$$

and LaSSI similarity between descriptor D_i and molecule

$$V_j = \sum_{x=1}^{k} \frac{Pix}{|Pi|} \cdot \frac{Qjx}{|Qj|}$$

The calculation of LaSSI similarity between an ad hoc query molecule v and the database molecule V_i is given by

$$Sim_{vVi} = \sum_{x=1}^{k} \frac{vx}{|v|} \cdot \frac{Qix}{|Qi|}$$

where the Qix are elements of Q_k with x ranging from 1 to k (total number of singular values). Sim_{vVi} ranges from −1.0 (least similar) to 1.0 (most similar).

Merging Textual and Chemical Information. Given the individual success of LSI[13] and LaSSI,[5,6] it seems natural to merge the textual and chemical descriptors (Figure 1). From

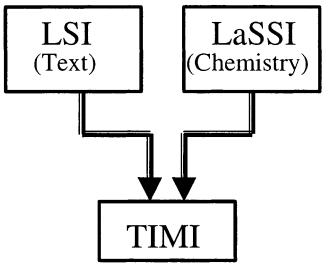


Figure 1. The schematic representation of the information domains embedded and explored by TIMI.

one perspective this could be perceived as adding more textual descriptors to the chemical descriptors representing each compound. From another perspective this is simply preserving the real-world context that one typically finds these compounds in when reading the scientific literature. In either case, the idea is to create a system whose whole is greater than the sum of its parts, i.e., to find associations between the text and chemical descriptors that could not be found by combining separate text and chemical analyses.

The approach we have taken is to add the chemical descriptors of compounds described in the scientific abstracts to the text in these abstracts, thereby enriching the text with chemistry. The following sections describe how this merging is done and show several retrieval and data mining scenarios using abstracts from the PubMed/MEDLINE (National Library of Medicine) resource.

METHODOLOGY

There are two phases of operation associated with TIMI. The first phase involves the creation of a TIMI database from a collection of abstracts or documents and the second phase involves querying that database. Almost all of the effort is expended during the first phase. After that, querying is very simple and consequently very fast.

Constructing a TIMI Database.

(1) Text Processing. The construction of a TIMI database begins with the collection of a set of documents in ASCII format. These documents might be journal articles, MEDLINE abstracts, internal progress reports, memos, trip reports, meeting minutes, etc. The native formats of these documents might require the use of conversion software to generate ASCII versions. The ASCII corpus is then normalized: unnecessary punctuation is removed, words are stemmed, case is normalized, and formatting is removed.

There are some idiosyncrasies of medical texts that make this step more challenging than it might be if we were analyzing texts from other fields. Systematic chemical names described in Chemical Abstracts[14] or International Union of Pure and Applied Chemistry (IUPAC)[15] nomenclatures may contain parentheses, brackets, commas, single quotes, colons, hyphens, plusses, and periods. Gene and protein names are often short acronyms which can be confused with other words when case has been normalized. Database identifiers and accession numbers can also obfuscate normalization. We use Perl scripts with access to specially crafted lexicons of chemical, gene, and protein names, and identifiers to perform the text processing necessary to normalize the input documents.

(2) Compound Identification. The terms of each normalized document are compared against an index of chemical compound names with known chemical structures provided by the Compound Knowledge Base (CKB) system.[16] One instance of each matched term is retained to create a list of unique structures for augmentation of the text.

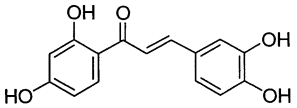


Figure 2. Butein.

Table 1. Partial List (10 out of the 56) of Descriptors of Butein and Their Term Frequencies

descriptor	no. of occurrences
c21c2101	3
c21c2102	4
c21c2103	6
c21c2104	4
c21c2105	2
c21c2106	2
c21c2107	5
c21c2108	2
c21c31c21c21	5
c21c31c31c21	3

(3) Descriptor Generation. A file of the connection tables of these structures is later used to generate their atom pair and topological torsion descriptors.[17,18] A textual representation of the chemical descriptors is then merged with the original text. For example, consider the following abstract title as a document.

"*Butein, a specific protein tyrosine kinase inhibitor*."

After normalization, this document would contain the seven words "butein", "a", "specific", "protein", "tyrosine", "kinase", and "inhibitor". The structure for butein, shown in Figure 2, exists in CKB.

The butein connection table generates 56 atom pair and topological torsion descriptors,[4,5] some of which are shown in Table 1. We can think of the descriptors as terms and merge them directly into the text of abstracts mentioning butein.

After this stage of the processing, the representation of the abstract title document would be the seven English words (each occurring once) and the 56 chemical terms (each with their own frequencies), for a total of 63 terms. Note that the word "a" still exists because stop word removal is yet to be performed.

A list of stop words are generated from inverse document frequency (idf) scores—any term occurring in more than 50% of the documents is removed from consideration as a row of the matrix. The merged text and chemistry is then recast to create a matrix where each row represents a unique term, each column represents a document, and the value of element $\langle i, j \rangle$ is the number of occurrences of term i in document j. An SVD of this matrix is performed resulting in the three SVD matrices used in calculating similarities.

TIMI Database Exploration. Searching a TIMI database is quite simple. The user specifies one or more words and/or chemical structures as a probe. Chemical structures are converted into chemical descriptors as was previously described. The frequency of occurrence of the words and the chemical descriptors are used to create a probe vector. This probe vector is projected into a lower dimensional space through the matrix multiplications described in the mathematics section. The user must select the dimensionality of

this space, i.e., the number of singular values, k, to use in the calculation. We have found that the optimal value of k depends on the nature and size of the database.[6]

Once it has been projected into the k-dimensional space, similarity between the probe and the documents and between the probe and the terms or descriptors can be computed as the cosine between their respective vectors. Documents and terms can then be sorted by their similarity to the probe and the user examines the top n ($n = 50$, 100, 200, etc.) documents or terms.

Database searching is accomplished by simply providing the ranked list of documents. Mining the database is a bit more interesting. TIMI was developed to assist pharmaceutical scientists in their efforts to retrieve information related to drugs, possibly discover new lead compounds, and to understand more about chemical structures and their relationships to the biological data mentioned in the literature. Therefore, we have investigated specialized mining tasks that can be addressed with TIMI including the extraction of chemical similarities and biological properties and associations.

For example, one can project one or more chemical structures into the k-dimensional space and then examine the list of compound identifiers that are similar. Both of these operations involve comparison between chemical structures although the similarity has been altered and perhaps enhanced by the presence of the surrounding text.

We can also calculate the similarity of a chemical probe to classes of terms in an effort to infer certain properties or relationships. We can examine the sorted list of terms to see what are the highest ranked therapeutic terms, disease names, toxic liabilities, adverse effects, etc. This is extremely important and to our knowledge a novel capability of TIMI. Suppose we determine that the rankings of therapeutic terms (terms related to therapeutic categories) heavily favor one category over all others. If we find that in the list of most similar terms to a particular compound are the words "cholesterol", "lipid", and "triglyceride", we might infer that there is some component of the compound's structure which is similar to the structures of compounds mentioned in abstracts about hypercholesterolemia. The same is true for highly ranked disease names or toxicity related terms such as "mutagen(ic)", "carcinogen(ic)", "hepatotoxic(ity)", etc.

Alternatively, we can look for associations from another perspective, that is, we can see which chemical descriptors are most similar to certain English terms. Consider the following question: which chemical descriptors are most associated with the terms "carcinogen" and "carcinogenic"? We create a probe vector with two nonzero frequencies corresponding to each term above. Then we examine the list of all terms ranked most similar to this probe looking for the highest ranked chemical descriptors. In practice, this type of query will result in a mixture of English and chemical terms that are related to the probe. In this case, we filter out the English terms so that only the chemical descriptors remain. The associated scores of these descriptors can be used to color the atoms of compounds of interest.[4,5] Coloring the atoms visually indicates which components of the compound are associated with the property. This approach can be taken to highlight any property that is described in the corpus.

Early identification of potential uses for and/or problems with new drugs can save pharmaceutical companies millions of dollars in research and development costs. TIMI allows the researcher to take advantage of past experiments described in the literature to address these concerns. We examine some of these relationships in the context of a corpus of MEDLINE abstracts in the next section.

MEDLINE ABSTRACTS: EXPERIMENTS AND RESULTS

MEDLINE Database. We used the PubMed/MEDLINE resource to access the medical research abstracts. A set of 11 571 MEDLINE abstracts using the term "drug" and published within a 3-month period of 1998 were extracted from the MEDLINE database.

MEDLINE abstracts, like all abstracts, typically contain only key findings, and hence we cannot expect to find relationships between compounds and textual terms that are found only in the full-length articles. However, we are quite confident, based on the results presented here, that the TIMI methodology would find the underlying concepts and relationships described in full-length articles if they were presented to it as input.

Two other databases were constructed in addition to the $TIMI_{TC}$ database described below to explore the information mining offered by each of them. A database of just the original textual terms, i.e., no chemistry, was created ($TIMI_T$), as was a LaSSI database containing just the chemical descriptors ($TIMI_C$).

database	contents
$TIMI_T$	textual terms
$TIMI_C$	chemical descriptors
$TIMI_{TC}$	textual terms and chemical descriptors

$TIMI_{TC}$ Database Characteristics. The text was preprocessed in order to identify chemical names and to merge the chemical descriptors of recognized compounds into the appropriate abstract(s). A total of 2876 unique compound identifiers whose connection tables exist within CKB were found within 6929 abstracts; 4642 abstracts did not have any identifiable structure associated with them. The 10 most frequently cited compounds were glutathione (181), dopamine (179), glucose (157), cholesterol (141), cisplatin (132), serotonin (131), cocaine (127), doxorubicin (111), adenosine (110), and morphine (109). The atom pair and topological torsion descriptors of all the compounds were added to the text. The list of chemical and textual descriptors was then used to create a term/abstract matrix. The dimensions of this matrix were 42 566 unique terms $\times$ 11 571 abstracts. The Lanczos iterative SVD algorithm[12] was used to produce 160 singular vectors in just under 7 min on one node of an IBM SP2 minicomputer. We will refer to this database as $TIMI_{TC}$.

Experiments. Three different sets of queries with 12 different examples were posed to the $TIMI_{TC}$ database (Table 2) to determine the number of singular values that yields the best retrieval of abstracts. The first set involved textual terms, the second set involved chemical structure queries, and the third set involved combined structure and textual queries.

The keywords, their biological activities, and the number of abstracts containing these keywords are given in Table 2.

Table 2. Keyword Queries and the Number of "Relevant" Abstracts Containing Them

keyword query	activity	no. of relevant abstracts[a]
ciprofloxacin	antiinfective	71
dopamine	dopamine receptor agonist	185
fluoxetine	antidepressant	33
glucose	nutrient	189
glutathione	detoxifier	193
indinavir	antiviral (HIV)	27
indomethacin	antiinflammatory	66
losartan	antihypertensive	18
simvastatin	anticholesterol	16
sumatriptan	antimigraine	7
tamoxifen	antiestrogen	50
troglitazone	antidiabetes	18

[a] Relevant abstracts are those that contain the query keyword at least once.

Table 3. Query: Keyword; Database: $TIMI_{TC}$—Initial Enhancements @100 and Number of Abstracts Retrieved from Top 100 Ranking Documents

keyword query	initial enhancement[a]	no. of abstracts retrieved	best *k*
ciprofloxacin	90	55	120, 130, 160
dopamine	56	89	150
fluoxetine	112	32	80−160
glucose	44	72	140
glutathione	60	100	70, 90−160
indinavir	111	26	20, 110−120
indomethacin	105	60	150 ,160
losartan	116	18	80, 100−160
simvastatin	116	16	50−160
sumatriptan	116	7	110, 120, 140−160
tamoxifen	111	48	140
troglitazone	116	18	130

[a] Initial enhancement = (*abstracts* @ 100/*nabstracts* × 100/11571), where *abstracts*@100 is the number of relevant abstracts retrieved in the top 100 scored abstracts and *nabstracts* is the total number of relevant abstracts given in Table 2. Initial enhancement, by the above definition, is a factor by which a retrieval performs over random retrieval.

Results. The keywords, structures of these keyword compounds, and the combination of keyword query and structure query were individually posed to the $TIMI_{TC}$ database and for each type of query *k* was varied from 10 to 160 in increments of 10. Each of the 11 571 abstracts was scored against each of these queries. We defined "relevant" abstracts as those that contain the query keyword at least once. Thus, the number of relevant abstracts containing the query keyword, listed in Table 2, retrieved in the top ranking 100 abstracts for each of these individual searches was used to calculate initial enhancements and the fraction of relevant abstracts retrieved. The results of these calculations are presented in Tables 3 and 4 and Figures 3−5.

The results of text queries against the $TIMI_{TC}$ database are given in Table 3. The initial enhancements, the most number of documents retrieved at the stated best *k* value(s)

$$\text{initial enhancement} = \frac{abstracts\ @100}{nabstracts \times 100/11\,571}$$

where *abstracts*@100 is the number of relevant abstracts retrieved in the top 100 scored abstracts and *nabstracts* is the total number of relevant abstracts given in Table 2. Initial

Table 4. Query: Structure; Database: $TIMI_{TC}$—Initial Enhancements @100 and Number of Abstracts Retrieved from Top 100 Ranking Documents

structure query	initial enhancement	no. of abstracts retrieved	best *k*
ciprofloxacin	108	66	150
dopamine	48	77	160
fluoxetine	116	33	130−160
glucose	39	63	160
glutathione	60	100	60−160
indinavir	111	26	30−160
indomethacin	105	60	160
losartan	116	18	100−160
simvastatin	116	16	80−160
sumatriptan	116	7	130−160
tamoxifen	106	46	90
troglitazone	109	17	90−160

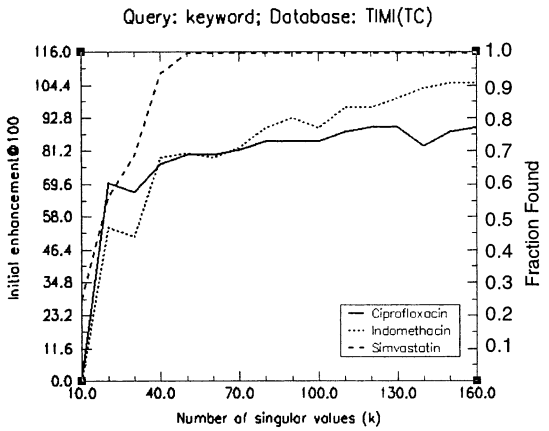


Figure 3. Query: keyword; database: $TIMI_{TC}$. Initial enhancements and fraction of relevant abstracts found for keyword searches of $TIMI_{TC}$ database. Initial enhancement is a measure of retrieval performance relative to random retrieval (see text for definition).

enhancement, by the above definition, is a factor by which a retrieval performs over random retrieval.

The results presented in Tables 3 and 4 show that searches of the $TIMI_{TC}$ database with chemical structures yielded results similar to those obtained with keyword searches against this database. This indicates that TIMI treats the keyword and its structure as synonyms in retrieving relevant documents to the front of the ranked documents. However, plots of the combined initial enhancement and fraction found (Figures 3−5) show that the keyword, structure, and joint keyword and structure queries, though they give similar results, yield different profiles. The indomethacin example presented below examines the differences in the results obtained with keyword, structure, and combined keyword and structure queries.

The goal of the above exercises was to determine the optimal parameters with which we can rank the maximum number of abstracts containing the keywords at the top of the list. If the goal of a literature search is to retrieve all documents that contain the keyword, then a straightforward keyword search should accomplish this. However, our hypothesis is that use of the TIMI system will allow retrieval of most of the documents containing the keyword and in addition those documents that contain the terms and compounds that are correlated to the keyword and/or to its structure. Thus by this virtue, we believe, the TIMI system

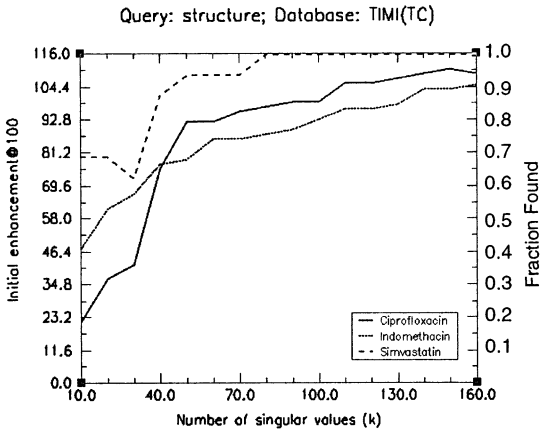


Figure 4. Query: structure; database: $TIMI_{TC}$. Initial enhancements and fraction of relevant abstracts found for chemical structure searches of $TIMI_{TC}$ database.

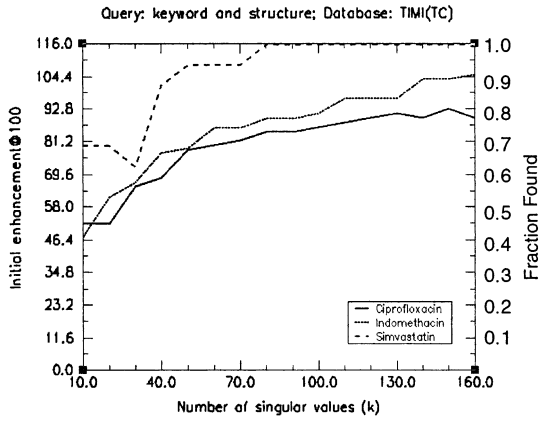


Figure 5. Query: keyword and structure; database: $TIMI_{TC}$. Initial enhancements and fraction of relevant abstracts found for combined keyword and chemical structure searches of $TIMI_{TC}$ database.

offers a distinct advantage over all other search facilities.

Although the results of the combined keyword and structure query are slightly different from independent keyword and structure queries, they all are qualitatively similar (Figures 3−5). However if we look closely in terms of the actual documents retrieved by each of the queries we can see they are in fact different.

Thus, to analyze the differences in the performance of each of these queries we look at the results for indomethacin, a widely prescribed antiinflammatory drug. It is worth noting that for all the three different sets of queries, with the current database, optimal document retrieval performance is achieved with $k = 130$ or higher. Therefore we use $k = 130$ in the indomethacin example.

Indomethacin Example. Three different queries were posed to the three TIMI databases. The first query is the indomethacin chemical structure, the second query is the textual term "indomethacin", and the third query is the combined structure and the name of indomethacin. Obviously, a structure query cannot be posed to the text only database and keyword query cannot be posed to the structure only database. The purpose of these three queries was to investigate the differences in retrieval and mining offered by each database.

Chemical Structure Query. We investigate in detail the performance of indomethacin (Figure 6) as a query against

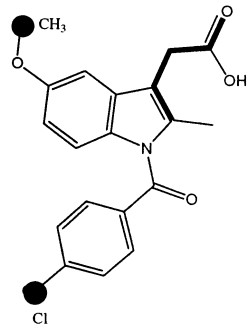


Figure 6. Indomethacin: The two overlapping topological torsion descriptors c31c31c20c31 and o11c31c20c31 are highlighted with thick bonds. The atom pair descriptor cl10c1012 is highlighted in ball rendering.

Table 5. Query: Indomethacin Structure; Database: $TIMI_{TC}$—Top Ten Scoring Abstracts and Terms at $k = 130$[a]

abstract	score	term	score
MED705	0.995	indomethacin	0.937
MED5516	0.993	TPDS	0.905
MED2632	0.990	gastropathy	0.879
MED4278	0.976	c31c31c20c31	0.860
MED9576	0.969	Karlovy	0.834
MED4057	0.964	gelation	0.829
MED5849	0.962	cross-linking	0.822
MED5850	0.962	cl10c1012	0.818
MED5942	0.952	dentatum	0.802
MED4379	0.929	o11c31c20c31	0.800

[a] MEDLINE abstract numbers represent the serial order in which they are present in the database.

$TIMI_{TC}$. Indomethacin is mentioned a total of 140 times in 66 different abstracts. A search of $TIMI_{TC}$ with the structure of indomethacin and setting $k = 130$ resulted in the lists of ranked documents and terms shown in Table 5.

The top 10 abstracts, MED705−MED4379, all contain information on indomethacin.

The top 10 terms can also tell us something about this compound. The term indomethacin is the highest ranked term which might not seem interesting at first, but recall that our probe was only the chemical descriptors from the structure of indomethacin and did not include the word "indomethacin". TPDS, trans phase delivery system, is a mechanism used to deliver indomethacin.[19] The third term, gastropathy, is an adverse effect that indomethacin exerts on the gastrointestinal tracts of patients who take this drug. The fourth term, c31c31c20c31, is a topological torsion chemical descriptor highlighted in Figure 6. Karlovy is a University in Czech Republic where studies with indomethacin were carried out. Ionotropic gelation was a technique used to formulate an oral drug delivery system for indomethacin.[20] The seventh term is about the cross-linking of indomethacin with glutaraldehyde to reduce swelling of guar gum in rats.[21] The term, cl10c1012, is an atom pair descriptor highlighted with ball rendering in Figure 6. Indomethacin was used in the inhibition of the migration of third-stage larvae of oesophagostomum dentatum.[22] O11c31c20c31 is a topological torsion, part of the acetic acid side chain of indomethacin highlighted in Figure 6.

We can perform the same query against the chemistry database $TIMI_C$. In this case we compute the LaSSI similarity of indomethacin to each of the other compounds found in

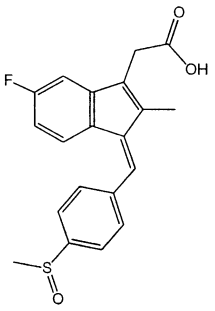
Table 6. Query: Indomethacin Structure; Database: $TIMI_C$—Top Ten Most Similar Compounds at $k = 130$

compound	no. docs	best rank	LaSSI similarity score
indomethacin	66	1	1.000
sulindac	6	30	0.612
clobazam	2	88	0.537
MK 886	2	131	0.508
oxyphenbutazone	1	46	0.485
aniracetam	3	132	0.476
dipyrone/metamizol	2	44	0.476
tolmetin	3	30	0.463
FO-349	2	69	0.448
KW-6002	1	158	0.446

the abstracts. We can then retrieve those articles which mention the high-ranking compounds. The top 10 ranking compounds and their similarity scores are given in Table 6.

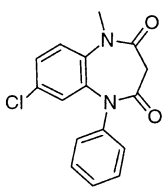
We find that many of the same compounds arising from abstracts selected from $TIMI_{TC}$ appear as the most similar compounds in the chemistry only search.

Sulindac is a nonsteroidal antiinflammatory agent used to treat rheumatoid arthritis and osteoarthritis. This compound is highly similar in structure to indomethacin, and thus, as expected, it works through the same mechanism of action.

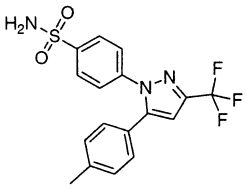


Sulindac

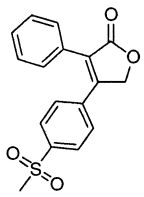
Clobazam, though it appears different from indomethacin at first glance, shares similar substructural features with indomethacin. Clobazam is an anticonvulsant used to treat epilepsy. The similarity of indomethacin to clobazam raises the issue of its possible CNS activity. A recent report indicates that COX-2 is expressed in the brain and its expression is increased in patients experiencing seizures.[23] This study also showed that indomethacin and celecoxib increased kainic acid-induced seizure and neuronal cell death in mice. This example illustrates how the structural similarity of indomethacin to clobazam embedded in the MEDLINE abstracts helped us identify the possible reason for its on-label adverse effects in epileptic patients. It is interesting to note that seizure is one of the side effects that appears as a warning on the label of the most recent COX-2 selective inhibitors, celecoxib and rofecoxib.



Clobazam

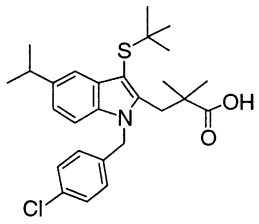


Celecoxib



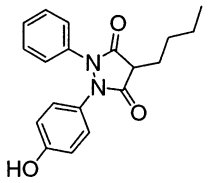
Rofecoxib

MK-886 is a 5-lipoxygenase activator protein inhibitor.[24]



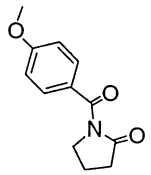
MK 886

Oxyphenbutazone, the fourth most similar compound, is used to treat chronic inflammatory conditions such as rheumatoid arthritis, osteoarthritis, and ankylosing spondylitis.[25] It is interesting to note that this compound is strikingly similar to the most recent COX-2 selective inhibitors, rofecoxib and celecoxib.



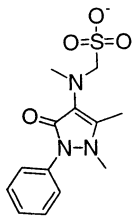
Oxyphenbutazone

Aniracetam is a cognition enhancing drug that works by modulating the function of AMPA receptors and protecting neuronal cells.[26]

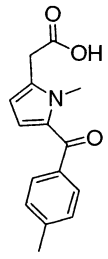


Aniracetam

Dypyrone is an over-the-counter antipyretic drug.[27]

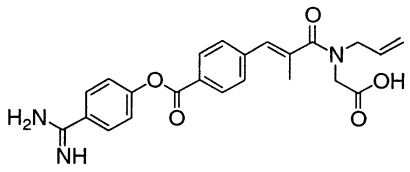


Dipyrone (Metamizol)



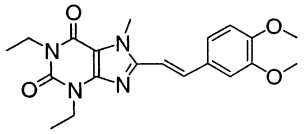
Tolmetin

Tolmetin is a prescription drug that is used to treat rheumatoid arthritis, osteoarthritis, and acute flares.[28]



FO-349

FO-349 exhibits antiproliferative activity via the inhibition of a serine protease.[29]



KW-6002

KW-6002 is undergoing late-stage clinical trials for its use to treat Parkinson's disease.[30] It exhibits adenosine A2A receptor antagonist activity. Indomethacin is known to cause aggravation of parkinsonism in patients suffering from this ailment.[28]

The results presented above show that four out of nine compounds in Table 6 work through a similar mechanism of action. The other five compounds reveal their interesting clinical/biological activities. The similarity of indomethacin to clobazam pointed to an on-label adverse effect that indomethacin exhibits in patients experiencing seizures. Indomethacin's similarity to KW-6002 pointed to its potential to aggravate symptoms in patients suffering from Parkinson's disease, though there is no published study to support this. The other nonantiinflammatory compounds, though similar to indomethacin, do not lead to abstracts that show indomethacin's interaction with their biological targets. However such information can be used to investigate the possible interaction of indomethacin with these targets and potentially avoid undesirable adverse effects. This example clearly demonstrates TIMI's unique ability to find embedded structure−activity relationships in the medical literature.

Text Query. If we use the term "indomethacin" to probe $TIMI_{TC}$ instead of the structure of indomethacin, we get results that are not that different from that of the structure query. The top 10 ranking abstracts and descriptors generated with the keyword indomethacin against the $TIMI_{TC}$ database are shown in Table 7. It is readily apparent that the abstracts and terms retrieved by the keyword query are very similar to the ones retrieved by the structure query (Table 5). This shows that indomethacin's name and its structure are synonyms in the $TIMI_{TC}$ database. This is consistent with the results presented in Tables 3 and 4. This observation is true in general, but the plot of correlation of ranks of the abstracts retrieved by keyword search versus those retrieved by the chemical structure search shown in Figure 7 reveals that there is a good degree of correlation for the most part but that quite a few abstracts that are ranked highly by the structure query are ranked quite low by the keyword query and vice versa. This is also true for the plot of correlation of ranks of abstracts retrieved by keyword query versus those retrieved by the combined keyword and structure query (Figure 8).

In the search of $TIMI_T$ with the keyword indomethacin, five out of 10 documents in Table 8 discuss indomethacin,

Table 7. Query: Text Term "Indomethacin"; Database: $TIMI_{TC}$—Top Ten Scoring Abstracts and Terms at $k = 130$

abstract	score	term	score
MED5516	0.884	indomethacin	0.980
MED705	0.880	gastropathy	0.927
MED2632	0.875	TPDS	0.926
MED4278	0.872	c31c31c20c31	0.869
MED9576	0.869	cl10c1012	0.862
MED4057	0.866	cross-linking	0.861
MED5850	0.862	Karlovy	0.851
MED5849	0.862	gelation	0.841
MED10520	0.856	patent	0.830
MED5942	0.840	dentatum	0.829

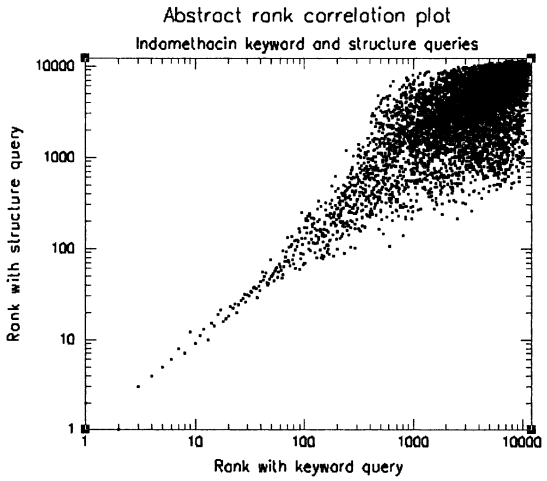


Figure 7. Correlation of abstract ranks for keyword "indomethacin" search and indomethacin structure search of the $TIMI_{TC}$ database.

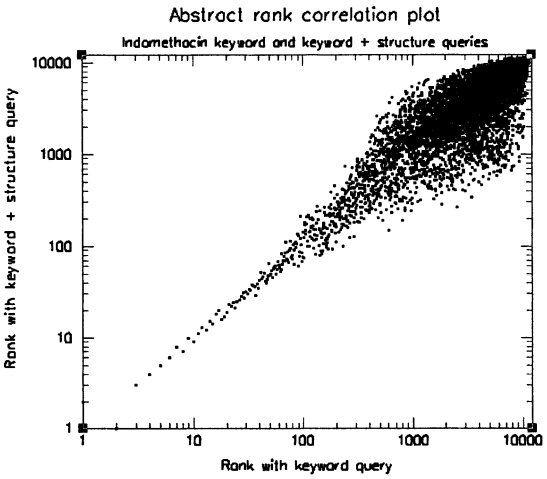


Figure 8. Correlation of abstract ranks for keyword "indomethacin" search versus combined keyword (indomethacin) and its chemical structure query of $TIMI_{TC}$ database.

but none of these abstracts were retrieved by the search using indomethacin structure or its name against $TIMI_{TC}$ (see Tables 5 and 7). The difference in the list of abstracts retrieved by the keyword query against $TIMI_T$ versus those retrieved by the same query against $TIMI_{TC}$ is most likely due to the lack of influence of the chemical descriptors. The first term, indomethacin, is found at the top of 10 terms as observed for the $TIMI_{TC}$ run, but after that they are all different. Clearly, the chemical descriptors are creating a qualitative difference in the rankings.

The plot of the correlation of ranks of keyword searches of the $TIMI_T$ database and $TIMI_{TC}$ database shows that there

Table 8. Query: Text Term "Indomethacin; Database: $TIMI_T$—Top Ten Scoring Documents and Terms at $k = 130$

document	score	term	score
MED4657	0.418	indomethacin	0.981
MED1010	0.377	cyclooxygenase	0.488
MED2330	0.335	prostaglandin	0.474
MED2328	0.335	tracheal	0.468
MED4379	0.332	antiinflammatory	0.466
MED5468	0.326	NSAID	0.460
MED8773	0.325	nonsteroidal	0.454
MED4942	0.324	Cracow	0.431
MED2548	0.315	rhabdomyolysis	0.431
MED7215	0.305	ulceration	0.426

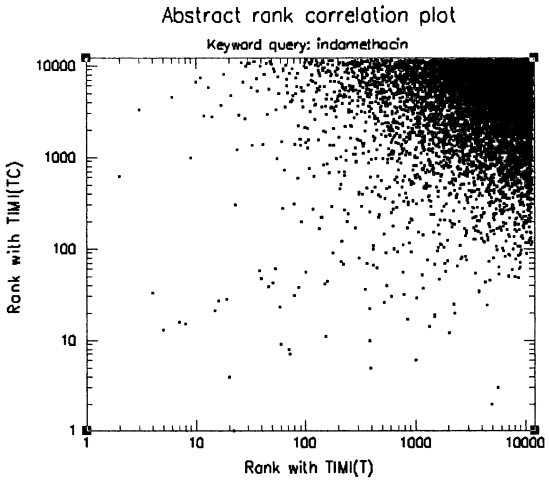


Figure 9. Correlation of abstract ranks for keyword "indomethacin" search of $TIMI_T$ database versus keyword search of $TIMI_{TC}$ database.

is very little correlation (Figure 9). This clearly demonstrates that the chemical descriptors in the $TIMI_{TC}$ database influence the ranks of the abstracts and bring completely different sets of articles to the top of the list compared to those retrieved by the keyword query of the text-only database.

Finally, the plot of abstract ranks generated by the indomethacin structure versus the word indomethacin and its chemical structure query against the $TIMI_{TC}$ database shows that it is highly correlated with a spindle-like structure centered along the diagonal (data not shown). This indicates that the large number of chemical descriptors representing the indomethacin structure is likely to dominate over the influence of the single keyword in scoring the abstracts.

DISCUSSION

Several interesting points arise from these experiments. The terms related to the structural query of indomethacin in the $TIMI_{TC}$ database are quite remarkable (see Table 6). The system uncovers associations between the chemical descriptors of the probe and many English words related to this antiinflammatory drug. The associations are along many different conceptual dimensions: the name of the probe, indomethacin; the chief biological activity, antiinflammatory; drug class, NSAID; chemical descriptors responsible for its activity, chlorine and methyl substituents and the carboxylic acid moiety; a specific mechanism of drug delivery; biological experiments; and affiliation information. There are other words whose rankings are not so obvious, and we believe that some of these terms might provide new insights.

The compounds found in highly ranked abstracts of the same search, sulindac, clobazam, MK-886, oxyphenbutazone, aniracetam, dipyrone, tolmetin, FO-349, and KW-6002 are interesting because four out of five compounds are from different biological activity categories. Specifically, the similarity of indomethacin to clobazam has revealed one of its adverse effects, seizure. It is less likely that a medicinal chemist interested in antiinflammatory drugs would know beforehand of this relationship uncovered by TIMI. This might be especially pertinent given the fact that this adverse effect appears on the label of the latest COX-2 inhibitors, celecoxib and rofecoxib (not present in the database presented here). Through this example we have shown TIMI's unique ability to reveal embedded structure/biological activity relationships in the medical literature. Uncovering these types of associations may allow researchers to discover novel pathways through which a compound can interact and thus provide opportunities to take advantage of them or find ways to avoid them.

CONCLUSIONS

The experiments above illustrate the advantages of merging textual and chemical descriptors over either text or chemistry individually. A text-only database cannot benefit from associations that are made across chemical structure; it cannot relate those textual terms to chemical features; and one can only retrieve documents about the compounds explicitly mentioned in the text. A chemistry-only database cannot benefit from associations made across the text nor can it index abstracts that do not have any chemical structures mentioned in them. The TIMI system represents a methodology for taking advantage of the contextual knowledge developed by scientists within the pharmaceutical, biological, and medicinal chemistry community.

ACKNOWLEDGMENT

We would like to thank all the members of Molecular Systems for their valuable comments.

REFERENCES AND NOTES

(1) Bevan, P.; Ryder, H.; Shaw, I.. Identifying small-molecule lead compounds: the screening approach to drug discovery. *Trends Biotech.* **1995**, *13*, 115−121.
(2) Willet, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.
(3) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley: New York, 1990.
(4) Hull, R. D.; Singh, S. B.; Nachbar, R. B.; Sheridan, R. P.; Kearsley, S. K.; Fluder, E. M. Latent Semantic Structure Indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.* **2001**, *44*, 1177−1184.
(5) Hull, R. D.; Fluder, E. M.; Singh, S. B.; Nachbar, R. B.; Kearsley, S. K.; Sheridan, R. P. Chemical Searches using Latent Semantic Structure Indexing (LaSSI). *J. Med. Chem*. **2001**, *44*, 1185−1191.
(6) Singh, S. B.; Sheridan, R. P.; Fluder, E. M.; Hull, R. D. Mining the Chemical Quarry with Joint Chemical Probes: An application of Latent Semantic Structure Indexing (LaSSI) and TOPOSIM (Dice) to chemical database mining. *J. Med. Chem.* **2001**, *44*, 1564−1575.
(7) Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391−407.
(8) Strang, G. *Linear algebra and its applications*; Harcourt Brace Jovanovich College Publishers: 1998; pp 442−452.
(9) Alter, O.; Brown, P. O.; Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *29*, 10101−10106.

(10) Gencer, N. G.; Tek, M. N. Electrical conductivity imaging via contactless measurements. *IEEE Trans. Med. Imaging* **1999**, *18*, 617−627.
(11) Tsurui, H.; Nishimura, H.; Hattori, S.; Hirose, S.; Okumura, K.; Shirai, T. Seven-color fluorescence imaging of tissue samples based on Fourier spectroscopy and singular value decomposition. *J. Histochem. Cytochem.* **2000**, *48*, 653−662.
(12) Berry, M.; Do, T.; O'Brien, G.; Krishna, V.; Varadhan, S. SVDPACKC (Version 1.0) User's Guide; UTK technical report CS-93-194, revised March 1996; Department of Computer Science, University of Tennessee: Knoxville, 1996.
(13) Berry, M. W.; Dumais, S. T.; O'Brien, G. W. Using linear algebra for intelligent information retrieval. *SIAM Rev.* **1995**, *37*, 573−595.
(14) Chemical Abstracts Service. *Naming and Indexing of Chemical Substances for Chemical Abstracts*; American Chemical Society: 1997.
(15) Panico, R.; Powell, W. H.; Richer, J. C. *A Guide to IUPAC Nomenclature of Organic Compounds (recommendations 1993)*; Blackwell Science: 1994.
(16) Walker, M. J.; Hull, R. D.; Singh, S. B. CKB − The Compound Knowledge Base: A text-based chemical search system. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1293−1295.
(17) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(18) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsions: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(19) Mikulak, S. A.; Vangsness, C. T.; Nimni, M. E. Transdermal delivery and accumulation of indomethacin in subcutaneous tissues in rats. *J. Pharm. Pharmacol.* **1998**, *50*, 153−158.
(20) Pillay, V.; Dangor, C.M.; Govender, T.; Moopanar, K. R.; Hurbans, N. Ionotropic gelation: encapsulation of indomethacin in calcium alginate gel discs. *J. Microencapsul.* **1998**, *15*, 215−226.
(21) Gliko-Kabir, I.; Yagen, B.; Penhasi, A.; Rubinstein, A. Low swelling, cross-linked guar and its potential use as colon-specific drug carrier. *Pharm. Res.* **1998**, *15*, 1019−1025.
(22) Daugschies, A.; Ruttkowski, B. Modulation of migration of Oesophagostomum dentatum larvae by inhibitors and products of eicosanoid metabolism. *Int. J. Parasitol.* **1998**, *28*, 355−362.
(23) Baik, E. J.; Kim, E. J.; Lee, S. H.; Moon, C. Cyclooxygenase-2 selective inhibitors aggravate kainic acid induced seizure and neuronal cell death in the hippocampus. *Brain Res.* **1999**, *843*, 118−129.
(24) Datta, K.; Biswal, S. S.; Xu, J.; Towndrow, K. M.; Feng, X.; Kehrer, J. P. A relationship between 5-lipoxygenase-activating protein and bcl-xL expression in murine pro-B lymphocytic FL5.12 cells. *J. Biol. Chem.* **1998**, *273*, 28163−28169.
(25) Loizzi, P.; Pipitone, V.; Bignamini, A. A. Double-blind clinical trial of protacine versus oxyphenbutazone in rheumatic disorders. *Pharmatherapeutica* **1980**, *2*, 285−292.
(26) Shirane, M.; Nakamura, K. Aniracetam enhances cortical dopamine and serotonin release via cholinergic and glutamatergic mechanisms in SHRSP. *Brain Res.* **2001**, *916*, 211−221.
(27) Wong, A.; Sibbald, A.; Ferrero, F.; Plage, R. M.; Santolaya, M. E.; Escobar, A. M.; Campos, S.; Barragan, S.; De Leon Gonzalez, M.; Kesselring, G. L. Antipyretic effects of dipyrone versus ibuprofen versus acetaminophen in children: results of a multinational, randomized, modified double-blind study. *Clin. Pediatr.* **2001**, *40*, 313−324.
(28) PDR electronic library. Copyright 2001 by Medical Economics Company Inc.: Montvale, NJ.
(29) Hiwasa, T.; Kondo, K.; Nakagawara, A.; Ohkoshi, M. Potent growth-suppressive activity of a serine protease inhibitor, ONO-3403, toward malignant human neuroblastoma cell lines. *Cancer Lett.* **1998**, *126*, 221−225.
(30) Knutsen, L. J.; Weiss, S. M. KW-6002 (Kyowa Hakko Kogyo). *Curr. Opin. Investig. Drugs* **2001**, *2*, 668−73.

CI025587A