Pacific Symposium on Biocomputing 12:304-315(2007)

MINING PATENTS USING MOLECULAR SIMILARITY SEARCH

JAMES RHODES[1], STEPHEN BOYER[1], JEFFREY KREULEN[1], YING CHEN[1],
PATRICIA ORDONEZ[2]

IBM, Almaden Services Research,
San Jose, CA 95120, USA
www.ibm.com
[1] *E-mail: jjrhodes, sboyer, kreulen, yingchen@us.ibm.com*
[2] *ordopa1@umbc.edu*

Text analytics is becoming an increasingly important tool used in biomedical research. While advances continue to be made in the core algorithms for entity identification and relation extraction, a need for practical applications of these technologies arises. We developed a system that allows users to explore the US Patent corpus using molecular information. The core of our system contains three main technologies: A high performing chemical annotator which identifies chemical terms and converts them to structures, a similarity search engine based on the emerging IUPAC International Chemical Identifier (InChI) standard, and a set of on demand data mining tools. By leveraging this technology we were able to rapidly identify and index $3,623,248$ unique chemical structures from $4,375,036$ US Patents and Patent Applications. Using this system a user may go to a web page, draw a molecule, search for related Intellectual Property (IP) and analyze the results. Our results prove that this is a far more effective way for identifying IP than traditional keyword based approaches.

Keywords: Chemical Similarity; Data Mining; Patents; Search Engine; InChI

1. Introduction

The US Patent corpus is an invaluable resource for any scientist with a need for prior art knowledge. Since patents need to clearly document all aspects of an invention, they contain an plethora of information. Unfortunately, much of this information is buried within pages upon pages of legal verbiage. Additionally, current search applications are designed around keyword queries which prove ineffective when searching for chemically related information.

Consider the drug discovery problem of finding a replacement molecule for fluoro alkane sulfonic acid ($CF_3CF_2SO_3H$). This molecule appears in everyday products like Scotchgard®, floor wax, Teflon®, and in electronic chip manufacturing materials like photo resists etc. The problem is that this molecule is a bioaccumulator and is a potential carcinogen (substance that causes cancer). Furthermore, it has made its way through the food chain, and can now be found in polar bears and penguins. Companies are pro actively trying to replace this acid with other more environmentally friendly molecules. The sulfonic acid fragment, SO_3H, is the critically necessary element. The harmful fragment is anything that looks like $CF_3(CF_2)_n$. The problem then is to find molecules that have the SO_3H fragment, and perhaps a benzene ring which would allow the synthetic chemist to replace an alkyl group with something that accounts for the electron withdrawing property of CF_3CF_2. The chemist would like to look for a candidate molecule based on its similarity to the molecular formula of the fragment, or the structure of the benzene or some weighted combination of both.

It is quite possible that the needed information exists in literature already, but may be costly and time consuming to discover. A system that would allow users to search and analyze documents, such as patents, at the molecular level could be a tremendously useful tool for biomedical research. In this paper we describe a system that leverages text mining techniques to annotate and index chemical entities, provide graphical document searching and discover biomedical/molecular relationships on demand. We prove the viability of such a system by indexing and analyzing the entire US Patent corpus from 1976-2005 and we present comparative results between molecular searching and traditional keyword based approaches.

2. Extracting Chemicals

The first step in the process is to extract chemical compounds from the Patent corpus. We developed two annotators which automatically parsed text and extracted potential chemical compounds. All of the potential chemicals were then fed through a name-to-structure program such as the name=struct®program from CambridgeSoft Corporation. Name=Struct makes no value judgments, focusing only on providing a structure that the name accurately describes.[1] The output of Name=Struct in our system is a connection table. Using the openly available InChI code,[10] these connection tables are converted into InChI strings.

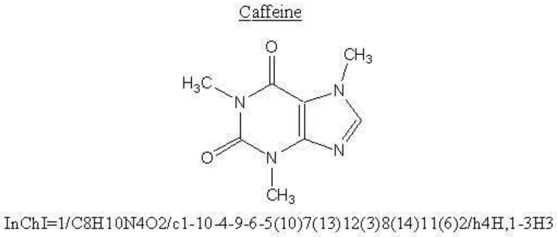
Due to the page limits, this paper focuses on the similarity search technology. We have built a machine learning and dictionary based chemical annotator that can extract chemical names out of text and convert them into structures. The similarity search capability is built on top of such annotation results, but is not tied to any specific underlying annotator implementation.

3. Indexing

As the new IUPAC International Chemical Identifier (InChI) standard continues to emerge, there is an increasing need to use the InChI codes beyond that of compound identification. Given our background in text analytics, we reduced the problem down to finding similar compounds based on the textual representation of the structure. Our experiments focused on the use of the InChI's as a method for identifying similar compounds. Using our annotators we were able to extract $3,623,248$ unique InChI's from the US Patent database (1976-2005) and Patent Applications (2001-2005). From this collection of InChI's an index was constructed using text mining techniques. We employed a traditional vector space model[14] as our underlying data structure.

3.1. *Vector Representation*

InChI's are unique for each molecule and they consist of multiple layers that describe different aspects of the molecule as depicted in Figure 1. The first three layers (formula, connection and hydrogen) are considered the main layers (see[15]) and are the layers we used for our experiments. Using the main layers, we extracted unique features from a collection of InChI codes.



InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Fig. 1. A compound and its InChI description

We defined features as one to three unique character phrases in the connection and hydrogen layers and unique atoms or symbols in the formula layer. Features from each layer are proceeded by a layer identifier. For the connection and hydrogen layers, features for an InChI i with characters c_j can be defined as unique terms c_j, $c_j{+}c_{j+1}$, $c_j{+}c_{j+1}{+}c_{j+2}$. These terms are added to the overall set of terms T which include unique c_j from the formula layer. Given a collection of InChI's I_i with terms T_j, each InChI is represented by the vector

$I_i = (d_{i1}, d_{i2}...d_{ij})$

where d_ij represents the frequency to the jth term in the InChI. For example, the two InChI's InChI=1/H2O/h1H2 and InChI=1/N2O/c1-2-3 would produce the following features

H, O, h1, h1H, h1H2, hH, hH2, h2, N, c1, c1-, c1-2, c-, c-2, c-2-, c2, c2-, c2-3, c-3, c3

with the following vector representations

$\{2,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0\}$ for water, and

$\{0,1,0,0,0,0,0,0,2,1,1,1,2,1,1,1,1,1,1\}$ for nitrous oxide.

In our experiments, the formula, connection and hydrogen layers produced 963, 69334 and 55256 features respectively. This makes the combined dimensionality of the dataset $T{=}125,553$. Feature values are always nonnegative integers. To take into account the frequency of features when computing the similarity distance calculation, we represented the vectors in unary notation where each of the three feature spaces is expanded by the maximum value of a feature in that space. This causes the dimensionality to exploded to $31,288,976$ features and the sparsity increases proportionally. Of course, this unary representation is implicit and need not be implemented explicitly.

Each InChI is processed by building for it three vectors which are then added to the respective vector space model. The results are three vector space models of size 309MB, 950MB and 503MB for the formula (F_1), connection (F_2) and hydrogen (F_3) layers.

Each vector space model F_j defines a distance function D_j by taking the Tanimoto[19] coefficient between the corresponding vectors. Consequently, for every two molecules x and y there are 3 distances defined between them, namely $D_1(x,y)$, $D_2(x,y)$ and $D_3(x,y)$.

3.2. *Index Implementation*

For indexing of the vector space models we implemented the Locality Sensitive Hashing (LSH) technique of Indyk and Motwani .[9] A major benefit of the algorithm is the relative size of the index compared to the overall vector space.

In our implementation the objects (and their feature vectors) do not need to be replicated. Vectors are computed for each InChI and stored only in a single repository. Each index maintains a selection of k vector positions and a standard hash function for producing an actual bucket numbers. The buckets themselves are individual files on the file system, and they contain pointers to (or serial numbers of) vectors in the aforementioned single repository. This allows both the entire index as well as each bucket to remain small. This implementation is of course useful because this single large repository still fits in our computer's main memory (RAM).

During index creation, not all hash buckets are populated. Additionally, the number of data points per hash bucket may also vary quite a bit. In our implementation, buckets were limited to a maximum of $B = 1000$. The end result is a LSH index L_j for each of the 3 layers of the InChI.

3.3. *Query Processing*

For each query molecule Q, vectors d_j are created from each vector space model F_j. Each vector is then processed by the LSH index L_j which corresponds to a given layer. The LSH index provides a list of potential candidate C_i which are then evaluated against the query vectors using the Tanimoto Coefficient. The total similarity for each candidate C_i is computed by

$$Si = \frac{\sum_{j=1}^{n} D(d_j, C_{ij})}{n} \tag{1}$$

where n is the total number of vector space models.

The Tanimoto Coefficient has been widely used as an effective measure of intermolecular similarity in both the clustering and searching of databases.[6] While Willet et al.[19] discuss six different coefficients for chemical similarity, we found that the Tanimoto Coefficient was the most widely recognized calculation with our users.

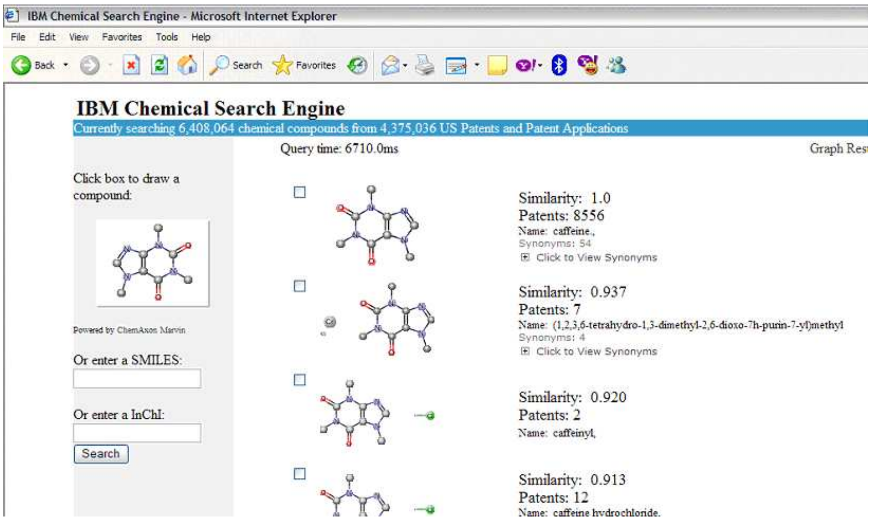
The results are then aggregated so each vector with the same S is merged into a set of synonyms. By dereferencing the vectors to the InChI's they represent and further dereferencing the InChI to the original text within the corpus, a list of the top K matching chemical names and the respective documents that contain those names is returned.

4. Experimental Results

In order to explain the experimental results, an overview of the application as it is currently implemented is required. We will conclude with a full description of the experimental process and its results.

4.1. *Graphical Similarity Search*

To use the Chemical Search Engine, a user may either draw the chemical structure of the molecule to be searched, enter an InChI or smile which represents the molecule into a text field, or open a file which stores a smile or InChI value in the corresponding field. The engine converts the query into an InChI and returns of a listing of molecules and their similarity values. Beside the molecule image is its similarity to the search molecule entered, its IUPAC name, an expandable list of synonyms, and the number of patents that were found containing that molecule as seen in Fig. 2. Not surprisingly for a query of a sketch of caffeine, the engine returned over 8,500 patents that contained a molecule with a similarity of 1.0, meaning that there was an exact match, and over 52 synonyms for that molecule. Six molecules with a similarity above .8 were rendered. For the experimental results, the canonical smile for the tested drug in the PubChem database was entered into the text field.



IBM Chemical Search Engine

Currently searching 6,408,064 chemical compounds from 4,375,036 US Patents and Patent Applications

Query time: 6710.0ms

Click box to draw a compound:

Powered by ChemAxon Marvin

Or enter a SMILES:

Or enter a InChI:

□ Similarity: 1.0
Patents: 8556
Name: caffeine.,
Synonyms: 54
⊞ Click to View Synonyms

□ Similarity: 0.937
Patents: 7
Name: (1,2,3,6-tetrahydro-1,3-dimethyl-2,6-dioxo-7h-purin-7-yl)methyl
Synonyms: 4
⊞ Click to View Synonyms

□ Similarity: 0.920
Patents: 2
Name: caffeinyl,

□ Similarity: 0.913
Patents: 12
Name: caffeine hydrochloride.

Fig. 2. Search results

4.2. *Molecular Networks*

In the upper right hand corner of the results page, the user may click on three different links to view selected molecules and their patents either as a graph using Graph Results, as a listing of hyper-linked patents with View Patents, or as an analysis of claims with Claim Analysis. In this section, we will describe and illustrate the usefulness of the Graph Results page and in the following, the Claim Analysis.

The value of a graphical representation of the selected molecules and their corresponding patents is most evident if we select the molecules with similar affinities to caffeine, but not exact matches to caffeine. The graph in Fig. 3 is a graph of the four molecules with the closest similarity to caffeine less than 1.

In the graph, the search node is fixed as the center node and molecular representations of the other nodes surround it. In the future, the graph will also display each molecule's similarity to the search node as indicted by the thickness of its edge to the center(search) node. When the user rolls over the center node, the comment "Search Node" is viewed whereas for the other nodes the name of the molecule is displayed. Note that some of the same molecules have different names.

The leaf nodes are the patents and patent applications associated with each molecule. If double-clicked the node will launch a browser window displaying the corresponding patent or application. A mouseover of these nodes will render the name of the assignee of the document. The nodes are color-coded by assignees.

A researcher may use this graph to view which molecules are most like the search node and of those molecules which have the greatest number of patents associated with them. It is also very useful for determining which assignees have the greatest number of patents for a particular molecular structure.

4.3. *Affinity Analysis*

The Claim Analysis page examines the claims of the patents associated with the selected molecules on the previous page to determine which medical conditions were found in the greatest number of patents. The more patents that mention a particular condition, the higher the condition's affinity to the molecule. Notice in Fig. 4, that for caffeine, migraine and headache have a high affinity, nausea and anxiety a moderate one, and burns and cough a low affinity.

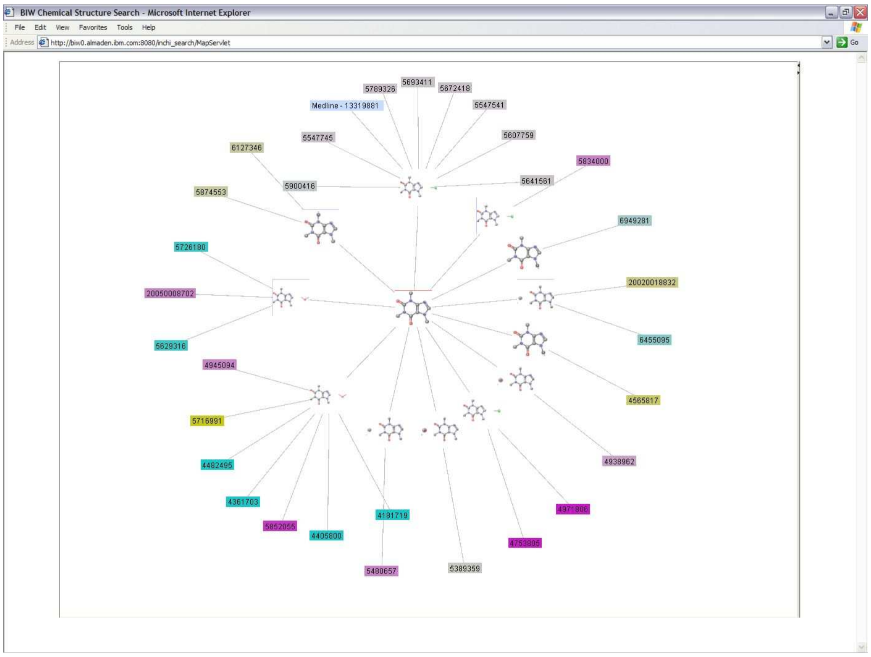


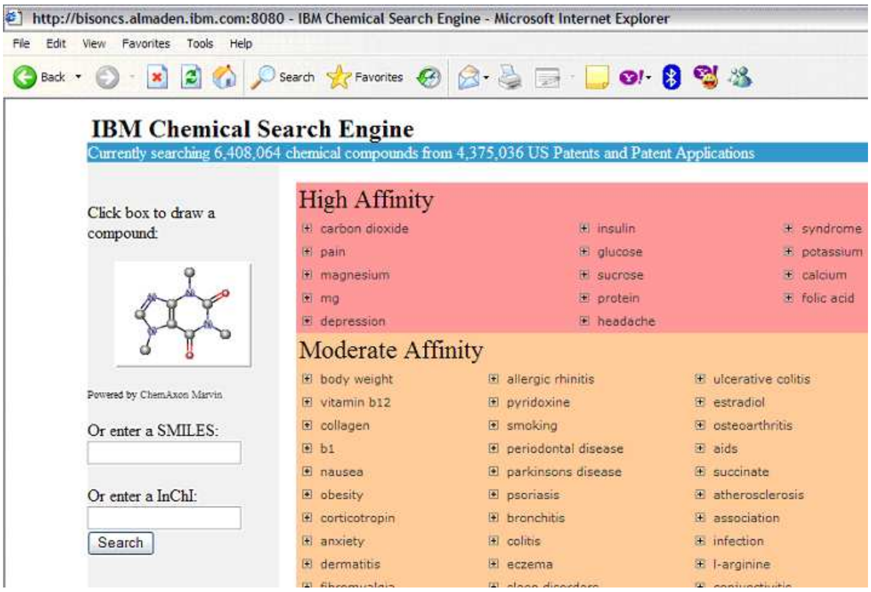
Fig. 3. Graph of selected molecules

The conditions were derived from a dictionary of proteins, diseases, and biomarkers. A dictionary based annotator annotates the full text of the selected patents in real time to extract the meaningful terms. A Chi-squared test was used referencing the number of patents that contained the conditions to determine the affinity between the molecules and the conditions.

On expanding a condition in the Claim Analysis page, a listing of the patents mentioning the condition in its text is rendered. The patent names are links to the actual patents. Thus, a researcher looking to patent a drug may do a search on the molecule and uncover what other uses the molecule has been patented for before. Such data may also serve to discover unexpected side effects or complications of a drug for the purposes of testing its safety.

4.4. *Results*

To evaluate the engine's effectiveness, we used a listing of the top 50 brand-name drugs prescribed in 2004 as provided by Humana.[8] We acquired a


http://bisoncs.almaden.ibm.com:8080 - IBM Chemical Search Engine - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Search

IBM Chemical Search Engine

Currently searching 6,408,064 chemical compounds from 4,375,036 US Patents and Patent Applications

Click box to draw a compound:

Powered by ChemAxon Marvin

Or enter a SMILES:

Or enter a InChI:

High Affinity

- ⊞ carbon dioxide
- ⊞ pain
- ⊞ magnesium
- ⊞ mg
- ⊞ depression
- ⊞ insulin
- ⊞ glucose
- ⊞ sucrose
- ⊞ protein
- ⊞ headache
- ⊞ syndrome
- ⊞ potassium
- ⊞ calcium
- ⊞ folic acid

Moderate Affinity

- ⊞ body weight
- ⊞ vitamin b12
- ⊞ collagen
- ⊞ b1
- ⊞ nausea
- ⊞ obesity
- ⊞ corticotropin
- ⊞ anxiety
- ⊞ dermatitis
- ⊞ allergic rhinitis
- ⊞ pyridoxine
- ⊞ smoking
- ⊞ periodontal disease
- ⊞ parkinsons disease
- ⊞ psoriasis
- ⊞ bronchitis
- ⊞ colitis
- ⊞ eczema
- ⊞ ulcerative colitis
- ⊞ estradiol
- ⊞ osteoarthritis
- ⊞ aids
- ⊞ succinate
- ⊞ atherosclerosis
- ⊞ association
- ⊞ infection
- ⊞ l-arginine

Fig. 4. Claims analysis of selected molecules

canonical smile value associated with each of the 25 top prescribed drugs from the PubChem database.[7] PubChem could not provide the smiles for two of the drugs, Yasmin 28 and OrthoEvra. If more than one molecule was returned from the database, we used the canonical smile value of the first one listed except in the case of three of the drugs, Tropol XL, Premarin, and Plavix. In these cases, we used the smile string that returned the greatest number of matches when we performed a search on the chemical search engine. With the generic name of the drug, we performed a search on one of the most sophisticated patent databases known, Delphion, using a boolean query that examined the abstracts, titles, claims, and descriptions of the patents for the name on patents from January 1, 1976 to December 31, 2005. The results can be seen in Fig. 5.

On acquiring the 25 drug names, the first obstacle was that 2 of the drugs could not be found in the PubChem database so that the canonical smile for these drugs could not be determined. Out of the 23 drugs that remained, our results indicate that for 19 of them more patents associated with the drug were found on our system than on Delphion. In the instances where the engine found more matches, the number of matches that it found was in some cases up to 10 times more, because the search was based on the molecular structure of the match and not on the generic name.

The number of times that a text based search outperformed the molecular search may be attributed a miss-selection of the smile string from the PubChem database. Thus, one of the greatest limitations of the chemical search engine is finding an accurate smile string for a given drug. Nevertheless, our experimental results demonstrate the enormous potential of being able to search the patent database based on a molecular structure.

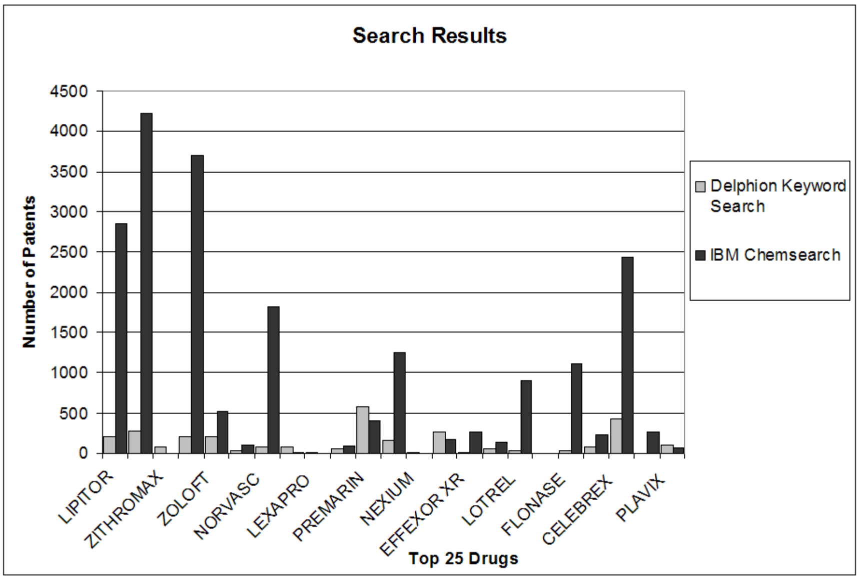


Fig. 5. A graph comparing the results of searching for the top 25 drugs listed by Humana[8] on the Chemical Search Engine using a molecular search and on DELPHION performing a text search of the compound's name.

5. Conclusion

We developed a practical system which leverages text analytics for indexing, searching and analyzing documents based on molecular information. Our results demonstrate that graphical structure search is a far more effective way to explore a document corpus than traditional keyword based queries when searching for biomedical related literature. The system is flexible and may be expanded to include other data sources besides Patents. These additional data sources would allow for meta-data information to be tied to Patents through chemical annotations. Future versions may allow researchers to explore data sets based on chemical properties such as toxicity or molecular weight. In addition to discovering literature for an exact match, this tool can be used for identifying practical applications of a compound or possible negative side effects by examining the literature surrounding similar compounds.

References

1. J. Brecher. Name=struct: A practical approach to the sorry state of real-life chemical nomenclature. *Journal of Chemical Information and Computer Science*, 39:943–950, 1999.
2. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gunshurst, D. L. Grier, B. A. Leland, and J. Laufer. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Science*, 32(3):244–255, 1992.
3. Inc. Daylight Chemical Information Systems. Daylight Theory: Fingerprints, 2005. `http://www.daylight.com/dayhtml/doc/theory/theory.finger.html`.
4. Inc. Daylight Chemical Information Systems. Daylight Cheminformatics SMILES, 2006. `http://daylight.com/smiles`.
5. GNU FDL. Open babel, 2006. `http://openbabel.sourceforge.net`.
6. D. Flower. On the properties of bit string-based measures of chemical similarity. *Journal of Chemical Information and Computer Science*, 38(3):379–386, 1998.
7. National Center for Biotechnology Information. Pubchem, 2006. `http://pubchem.ncbi.nlm.nih.gov/search`.
8. Humana. Top 50 brand-name drugs prescribed, 2005. `http://apps.humana.com/prescription_benefits_and_services/includes/Top50BrandDrugs.pdf`.
9. P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pages 604–613, may 1998.
10. IUPAC. The IUPAC International Chemical Identifier(InChI TM), 2005. `http://www.iupac.org/inchi`.
11. Stefan Kramer, Luc De Raedt, and Christoph Helma. Molecular feature mining in HIV data. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, 2001.
12. Elsevier MDL. Ctfile formats, 2005. `http://www.mdl.com/downloads/public/ctfile/ctfile.pdf`.
13. Elsevier MDL. Mdl isis/base, 2006. `http://www.mdli.com/support/knowledgebase/faqs/faq_ib_22.jsp`.
14. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

15. S. E. Stein, S. R. Heller, and D. Tchekhovskoi. An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier. In *Proceedings of the 2003 International Chemical Information Conference (Nimes)*, 2003.
16. Murray-Rust Research Group The University of Cambridge. The Unofficial InChI FAQ, 2006. `http://wwmm.ch.cam.ac.uk/inchifaq/`.
17. D. Weininger. Smiles, a chemical language and information system. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Science*, 28(1):31–36, 1988.
18. D. Weininger, A. Weininger, and J. L. Weininger. Smiles algorithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Science*, 29(2):97–101, 1989.
19. P. Willett, J. M. Barnard, and G. M. Downs. Chemical Similarity Searching. *Journal of Chemical Information and Computer Science*, 38(6):983–996, 1998.