Chemical Compound Navigator: A Web-Based Chem-BLAST, Chemical Taxonomy-Based Search Engine for Browsing Compounds

M.D. Prasanna,[1] Jiri Vondrasek,[2] Alexander Wlodawer,[3] H. Rodriguez,[1] and T. N. Bhat[1]*

[1] *Biochemical Science Division (831), NIST, Gaithersburg, Maryland*
[2] *Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Prague, Czech Republic*
[3] *Macromolecular Crystallography Laboratory, National Cancer Institute, Frederick, Maryland*

ABSTRACT **A novel technique to annotate, query, and analyze chemical compounds has been developed and is illustrated by using the inhibitor data on HIV protease-inhibitor complexes. In this method, all chemical compounds are annotated in terms of standard chemical structural fragments. These standard fragments are defined by using criteria, such as chemical classification; structural, chemical, or functional groups; and commercial, scientific or common names or synonyms. These fragments are then organized into a data tree based on their chemical substructures. Search engines have been developed to use this data tree to enable query on inhibitors of HIV protease (http://xpdb.nist.gov/hivsdb/hivsdb.html). These search engines use a new novel technique, Chemical Block Layered Alignment of Substructure Technique (Chem-BLAST) to search on the fragments of an inhibitor to look for its chemical structural neighbors. This novel technique to annotate and query compounds lays the foundation for the use of the Semantic Web concept on chemical compounds to allow end users to group, sort, and search structural neighbors accurately and efficiently. During annotation, it enables the attachment of "meaning" (i.e., semantics) to data in a manner that far exceeds the current practice of associating "metadata" with data by creating a knowledge base (or ontology) associated with compounds. Intended users of the technique are the research community and pharmaceutical industry, for which it will provide a new tool to better identify novel chemical structural neighbors to aid drug discovery. Proteins 2006;63:907–917.**

Key words: Semantic Web, Chem-BLAST, HIV, AIDS, chemical ontology, chemical taxonomy chemical knowledge base, InChI, drug discovery, mix and match method for drug design, chemical structural neighbor, substructure search, chemical data-tree

INTRODUCTION

Chemical databases play a major role in holding, annotating, and distributing chemical, biological, medicinal, and structural data (Enhanced NCI Database Browser: http://cactus.nci.nih.gov/; Chemistry WebBook: http://webbook.nist.gov/chemistry/; Thermodynamics Research Center: http://www.trc.nist.gov; Mass Spectral Library: http://www.nist.gov/srd/; Protein Data Bank: http://www.pdb.nist.gov; Relibase: http://relibase.ebi.ac.uk/reli-cgi/rll?/reli-cgi/general_layout.pl?home). These Web sites are an important part of modern means of information exchange on chemical structures (compounds). One of the major advantages of these Web-based resources is their federated approach to data standards and annotation. With use of such databases, vendors can provide the most up-to-date data with a single resource, and customers can get reliable access no matter where they are. In this global platform of Internet, simple but intuitive data annotation and navigation systems capable of producing complete and yet manageable results from a query are important issues.[1] Despite the wide availability and use of physical, chemical, and biochemical databases on the Web,[2,3] the ability to organize and retrieve structure-based data is challenging. Although it is possible to readily find compounds whose structural identifier (three-letter code[4] or InChI[5]) are known in advance, the ability of a user or of an automated search method to find similar substances in large, complex structural collections is, more often than not, unsatisfactory. Such searching or browsing serves at least two purposes: 1) to find the most closely related information[6] when data for a specific substance are not available and 2) to enable users to discover compounds with desired structural characteristics[7] for comparing existing drugs and for designing new ones. This is a problem for many users of major data collections, such as the Chemistry WebBook, the Protein Data Bank (PDB), and various databases developed at the Thermodynamics Research Center (TRC), where "hit lists" using conventional similarity or substructure search criteria often yield

Certain trade and company products are identified in this article to specify adequately the computer products needed to develop this data system. In no case, does such identification imply endorsement by the National Institute of Standards and Technology (NIST), nor does it imply that the products are necessarily the best available for the purpose.

*Correspondence to: T.N. Bhat, Biochemical Science Division (831), NIST, 100 Bureau Drive, Gaithersburg, MD 20899-8314. E-mail: bhat@nist.gov

Received 22 July 2005; Accepted 22 November 2005

Published online 28 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20914

large numbers of irrelevant retrievals and miss substances that the user wantedd to find.

Some data resources (e.g., the PDB[2]: http://www.pdb.nist.gov; SwisProt[8]: http://us.expasy.org/sprot/) provide query on chemical compounds using IUPAC names (ACD/ChemSketch[9]: http://www.acdlabs.com/products/chem_dsn_lab/chemsketch/tech.html). These text-based queries have limited capabilities because IUPAC names may not provide a unique index[10] for compounds. Moreover, users are accustomed to recognize compounds using molecular sketches rather than text strings. IUPAC names are also long and hard to memorize to be typed into a Web input box. HIV protease inhibitors are perfect examples of difficult cases to be specified and compared by using IUPAC names.

The principal difficulty in searching on structural fragments (fragments) of a compound is that structural features that are of interest to a user often cannot be defined (and indexed) in advance because of the natural complexity of structure/property relations,[11] which can depend on discipline, task, and user. Here we present an adaptive, customizable, automated method of processing and presenting fragments or substructures of compounds (connection tables) that are sufficiently flexible and easy-to-use and allow users to find, with confidence, information for the most structurally relevant fragments[7] in a data collection. This annotation enables the attachment of "meaning" (i.e., semantics) to substructures in a manner that far exceeds the current practice of associating "metadata" with data. This is accomplished by creating a knowledge base (or ontology) associated with each structure. During data annotation, compounds are mapped to standard fragments that may have chemical or functional meaning.[12] By using these fragments, compounds are then organized in a relational data tree[13] on the basis of their chemical substructure and semantics. This data tree is used by search engines to present compounds in layers of increasing granularity.[9] These search engines use the Semantic Web concept that aims for a more intelligent online experience where Web servers are written to be more intuitive and accurate in processing data and finding results for end users. These search engines are "data aware": they know from where a particular Web page came and where they can next lead to; they use this knowledge to guide users to formulate complex queries on fragments. Users specify chemical fragments to these search engines by using drop-down lists, text strings, or molecular sketches shown as hyperlinks. These queries are specified in several layers; in each layer, users refine the query element (probe) further to reduce unwanted hits.

Data Acquisition and Processing

Key components of creating an easy-to-use data resource are the efficient capture, context-based annotation, and layered distribution of data. In our case, data capture and annotation consists of acquiring the structural data from public data resources, format conversions, removal of redundant, inconsistent information, followed by data annotation with emphasis on data query and distribution. The data annotation step validates the data and makes the changes needed to transform the data into a common framework of data definitions and standards. The annotation step focuses also on customer needs during the query of structures by organizing data in several layers of low to high granularity. During a query, each layer of data is used for making a decision for subsequent queries.

In our method, data (atom names, chemical compound names, and three-dimensional atomic coordinates) are obtained from the PDB and from the HIVdb.[5,14] Each chemical compound is added into layer 4 and annotated in five steps (Fig. 1, annotation) to generate layers 5–1. In step 1, compounds from layer 4 (such as A77003, KNI-272) are broken into smaller fragments (layer 3) by using definitions provided in two dictionaries, one for the bonds (such as Cα-C…N-Cα) to be cleaved to generate fragments and the other for the names of the fragments (such as Phe and Thio-Proline) produced by the cleavage. A full description of the method used to define and break bonds will be published elsewhere. In step 2, these fragments are classified (layer 2) into standard groups (such as phenyl, thiazolidine) based on their substructures. In step 3, classes (such as six member rings, five member ring) are assigned (layer 1) to these groups. In step 4, data from steps 1 to 4 are organized into a data tree (Fig. 2). A data tree is a database table with one column for each layer of data (such as six-member ring → phenyl → Phe → A77003, space group) and one row for each unique compound, such as A77003, KNI-272. Information, such as synonyms of compounds, names of files with 2D static pictures of the substructures, PDBID, and crystallographic and biological data, is stored in columns 5 and above. For simplicity, in our application, we decided to use three layers to define substructures; however, the method would work as well with additional or fewer layers.

If the compounds are equated to proteins, then step 1 breaks them into amino acids and steps 2 and 3 define features, such as aromatic rings and other chemical and structural properties of these amino acids. In our application for peptidomimetic inhibitors, bonds are broken at or near the peptide bonds.[7] For nonpeptidic inhibitors, bonds are broken by considering both their functional similarity to fragments of peptidic inhibitors and the suitability of the resulting fragments to generate standard groups and classes in steps 2 and 3, respectively. Standard fragments (fragments, groups, classes) could have also been generated on the basis of other rules (e.g., RECAP procedure[15] or maximal block size[16]) or by invoking structural templates that are specific to the HIV protease inhibitors (e.g., a diol[17] or a norstatine moiety[18]). For indexing purposes, standard groups and fragments are assigned IUPAC names, an IUPAC International Chemical Identifier (InChI[19, 5]: http://www.iupac.org/projects/2000/2000-025-1-800.html), and 2D pictures of substructures and structures. The annotation steps also convert complicated connectivity tables of all structures and their substructures into text-based names. Therefore, general purpose text-based commercial database software, such as Oracle or MySQL, could be used to store, index, query, and

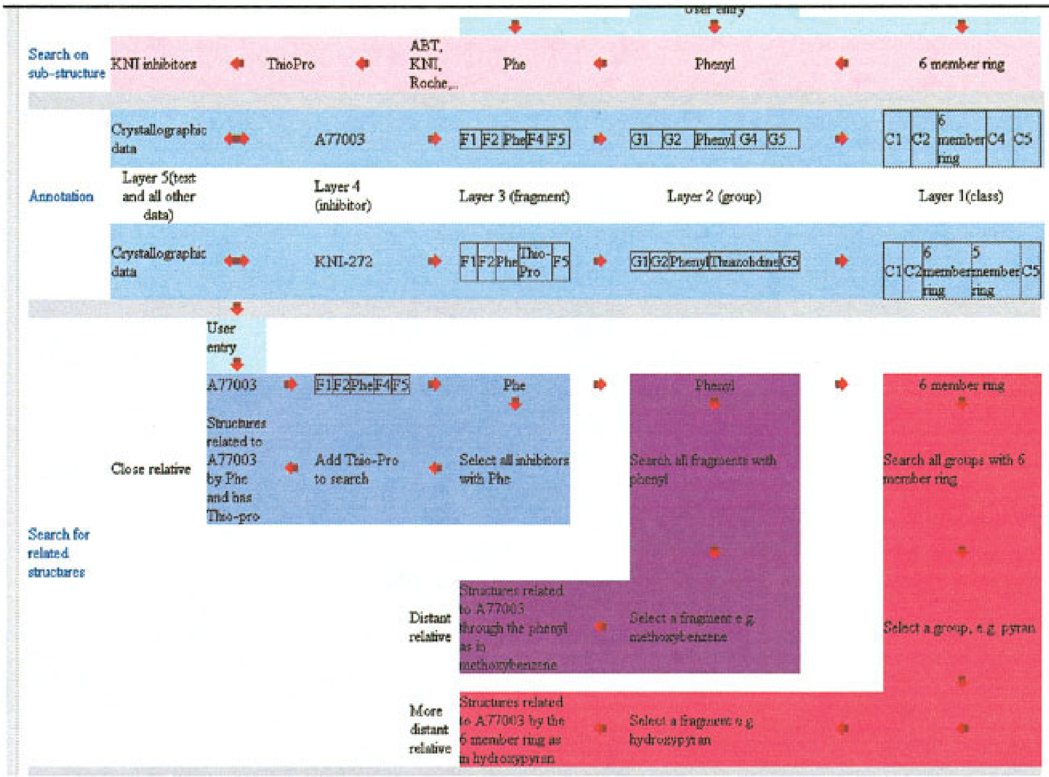

User entry
Search on sub-structure
KNI inhibitors
ThioPro
ABT, KNI, Roche,..
Phe
Phenyl
6 member ring
Crystallographic data
A77003
F1 F2 Phe F4 F5
G1 G2 Phenyl G4 G5
C1 C2 6 member ring C4 C5
Annotation
Layer 5(text and all other data)
Layer 4 (inhibitor)
Layer 3 (fragment)
Layer 2 (group)
Layer 1(class)
Crystallographic data
KNI-272
F1 F2 Phe Thio-Pro F5
G1 G2 Phenyl Thiazolidine G5
C1 C2 6 member ring 5 member ring C5
User entry
A77003
F1 F2 Phe F4 F5
Phe
Phenyl
6 member ring
Close relative
Structures related to A77003 by Phe and has Thio-pro
Add Thio-Pro to search
Select all inhibitors with Phe
Search all fragments with phenyl
Search all groups with 6 member ring
Search for related structures
Distant relative
Structures related to A77003 through the phenyl as in methoxybenzene
Select a fragment e.g. methoxybenzene
Select a group, e.g. pyran
More distant relative
Structures related to A77003 by the 6 member ring as in hydroxypyran
Select a fragment e.g. hydroxypyran

Fig. 1. The key features of annotation and search techniques proposed here. In the annotation step, the compounds are organized into a data tree of several layers. Layer 5 and higher have general crystallographic, 2D pictures, and text information on a compound. Layer 4 contains structural information on a complete compound. Layer 3 contains names of fragments of a compound. Layer 2 has names of the groups, and layer 1 has the names of chemical classes of the groups. Search on substructure of compounds is done by querying on columns containing either fragments or groups or classes. A list of structural neighbors of a particular structure is obtained by searching on one or more substructures of that structure. Arrows mark entry points to search engines and users.

illustrate structural data that are otherwise described by complicated connection tables of atomic distances and angles and that are therefore amenable only to special structural tools.

Search Engines

The arrangement of compounds into a data tree facilitates the development of a variety of search engines for querying the data. In our application, these search engines are mainly of two types: type I, which traverses from lower to higher layers along a data tree, and type II, which cycles between certain layers of a data tree. Search engine type I "zooms in" into the structural details of a compound; it produces hyperlinks to query on n^{th} layer to display structure from $(n + 1)^{th}$ layer where n can be 1, 2, 3, or 4. By contrast, search engine type II temporarily "zooms out" of structural details and "zooms back in" to focus on a specific section of the data tree. This search engine provides hyperlinks to query on a substructure in $(N - 1)^{th}$ layer and displays a substructure from N^{th} and $(N - 1)^{th}$ layer where N can be 2, 3, or 4. On choosing a hyperlink, it makes a query and displays the new results from both these layers. Among the new results, once again it displays the hyperlinks from $(N - 1)^{th}$ layer as earlier except that this time it appends the previous probe(s) to each one of these new hyperlinks. If a user clicks a fragment f1 in the first query and then he clicks a hyperlink for fragment f2 in the second query, then in the second query, this search engine uses both f1 and f2 for query. Search engine type II continues to operate between the same values of N with the concatenation of the probes between successive queries. In contrast, search engine type I moves up in n, but it ignores the previous probes (previous probes are subset of subsequent probes). Because search engine type I operates in successive layers, by the definition of a data tree all previous probes are structural subsets of the current probe (e.g., a six-member ring vs phenyl). Search engine type II is used in cases where the data for a given compound in a given layer has more than one unique fragment (such as Phe and Thio-Proline for KNI-272), and search engine type I is used for propogation of query from one to another layer of a data tree. Search engine type II performs a homology search of a structure displayed on the Web at a given time, and in successive steps, it allows a user to append additional fragments to a probe to define more closely related structural neighbors. Annotation steps 1–4 predefine the

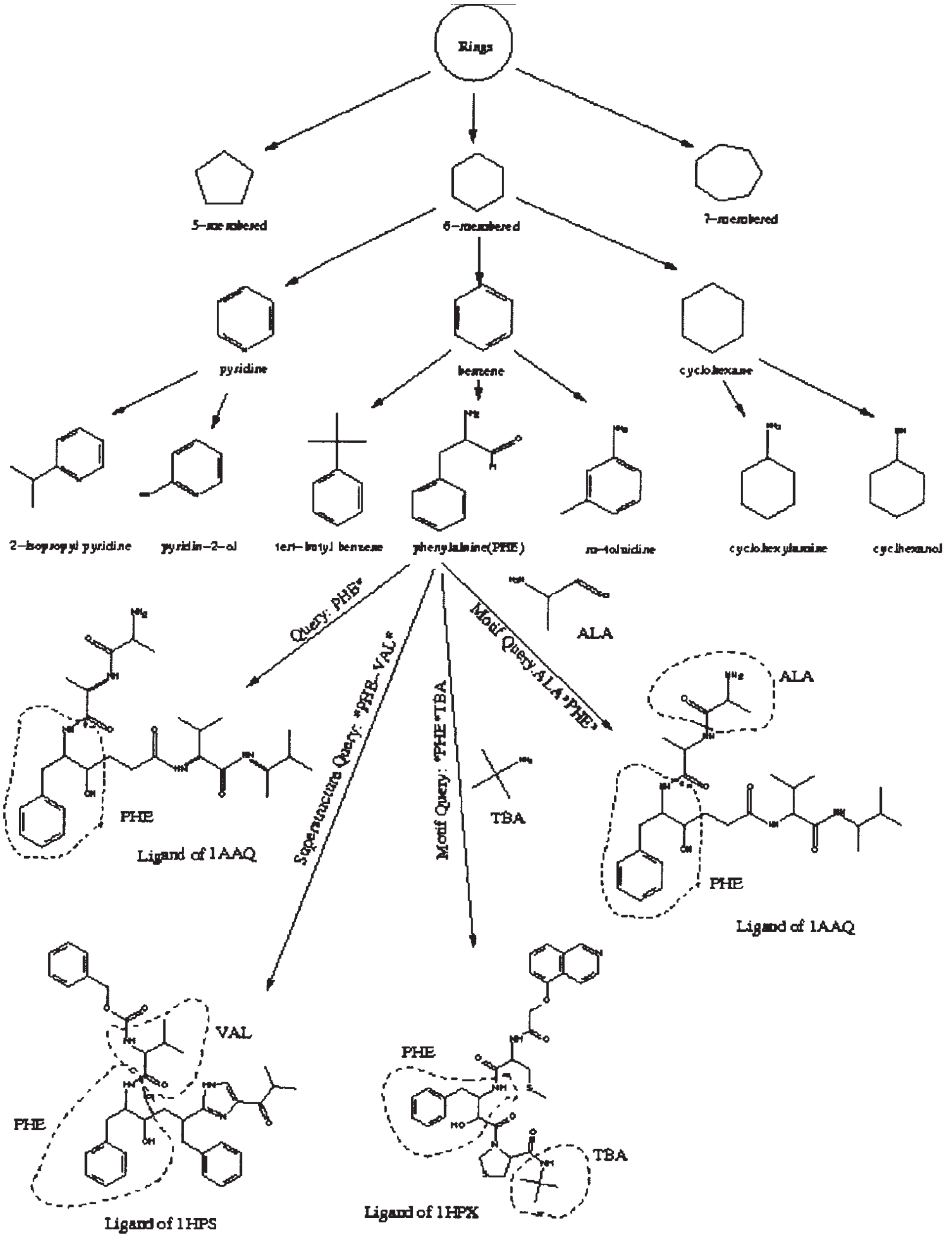
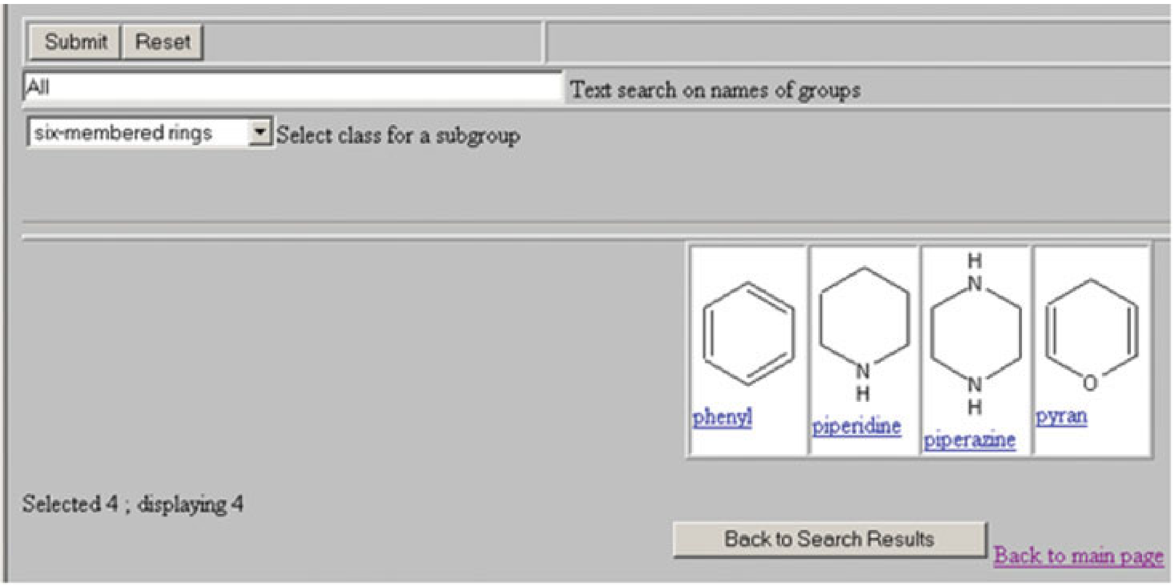


Fig. 2. The compounds are organized into a chemical data tree that places chemical structures into several layers. Search engines use these layers to present the structural data in distinct echelons to allow a user to define a probe of his interest. The search engines use this probe to perform Chem-BLAST and produce a list of chemical structural neighbors.

substructures of compounds; therefore, search engines are "data aware" (i.e., they are able to present the substructures of a given structure for a user to pick and choose for homology searches). The two types of search engines described above may be combined by the use of different values of N and n to generate several other search engines, which allow a user to crisscross between layers of the data tree and to experiment with different ways of defining structural neighbors to suit their needs. The actual number of search engines one may deploy in a given application depends on the number of layers in a data tree, user's interest, and the complexity of the data.

Fig. 3. The result of a query with search engine type I for six-member rings. The scroll bars for query are populated from layer 1, and the results are displayed from layer 2 where layer 1 has the value—six-member ring. (http://xpdb.nist.gov/hivsdb/hiv_ligands_class_to_subgroup.pl?T1=six-membered%20rings).

Fig. 4. The result of a query with search engine type I for fragments that contain phenyl ring. This query is done by the hyperlink on phenyl shown in Figure 3. For brevity, only a part of the result page is shown. (http://xpdb.nist.gov/hivsdb/hiv_ligands.pl?S1=phenyl).



All Text search on names of groups

six-membered rings Select class for a subgroup

Selected 4 ; displaying 4

Figure 3.

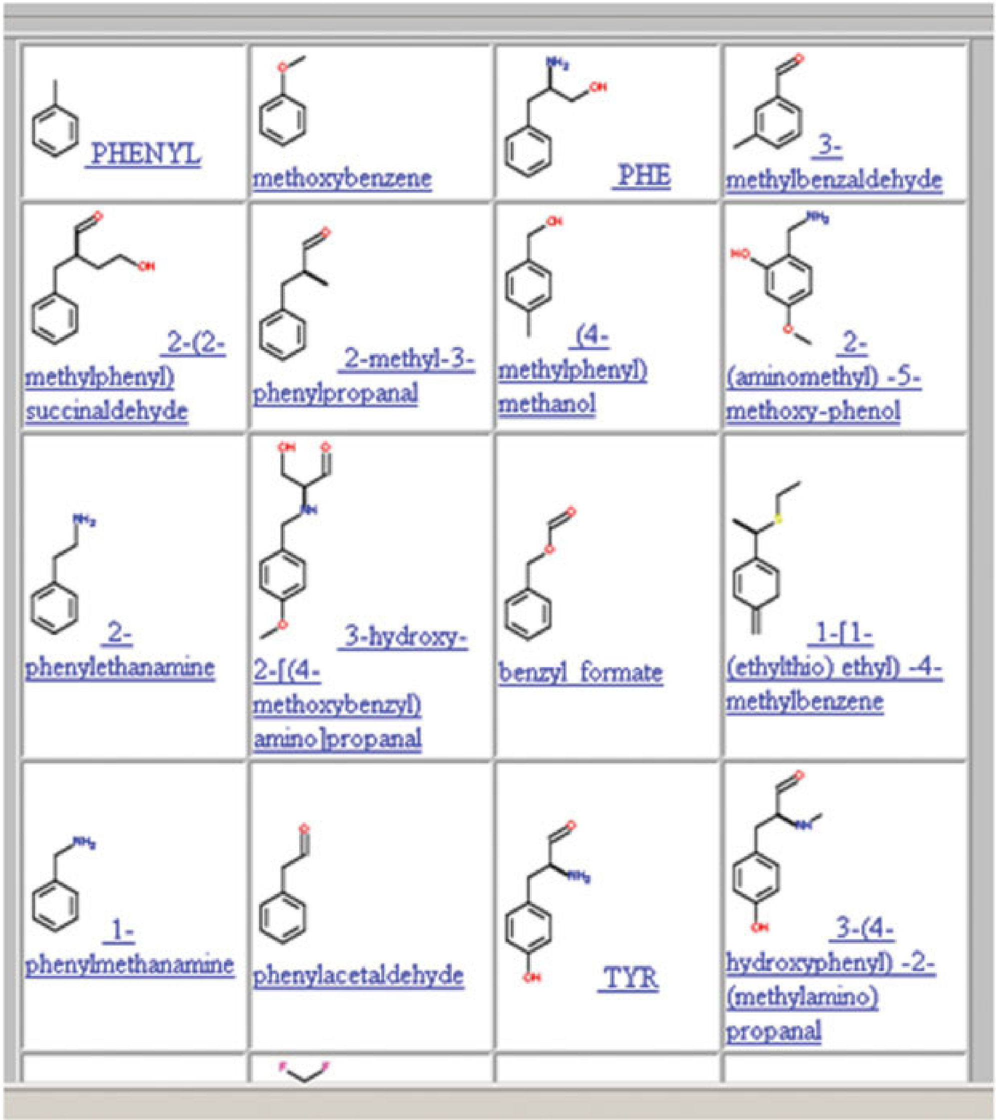
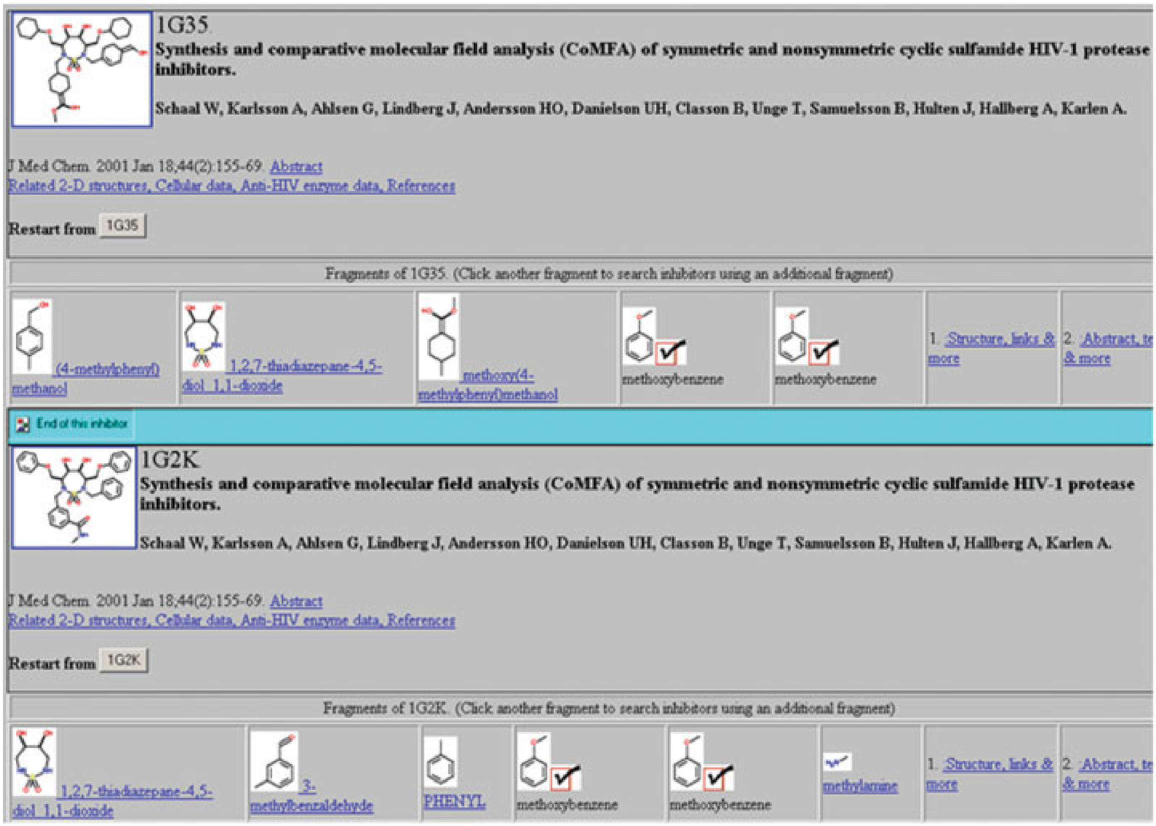

PHENYL
methoxybenzene
PHE
3-methylbenzaldehyde
2-(2-methylphenyl) succinaldehyde
2-methyl-3-phenylpropanal
(4-methylphenyl) methanol
2-(aminomethyl) -5-methoxy-phenol
2-phenylethanamine
3-hydroxy-2-[(4-methoxybenzyl) amino]propanal
benzyl formate
1-[1-(ethylthio) ethyl) -4-methylbenzene
1-phenylmethanamine
phenylacetaldehyde
TYR
3-(4-hydroxyphenyl) -2-(methylamino) propanal

Figure 4.



1G35
Synthesis and comparative molecular field analysis (CoMFA) of symmetric and nonsymmetric cyclic sulfamide HIV-1 protease inhibitors.

Schaal W, Karlsson A, Ahlsen G, Lindberg J, Andersson HO, Danielson UH, Classon B, Unge T, Samuelsson B, Hulten J, Hallberg A, Karlen A.

J Med Chem. 2001 Jan 18;44(2):155-69. Abstract
Related 2-D structures, Cellular data, Anti-HIV enzyme data, References

Restart from 1G35

Fragments of 1G35. (Click another fragment to search inhibitors using an additional fragment)

(4-methylphenyl) methanol	1,2,7-thiadiazepane-4,5-diol 1,1-dioxide	methoxy(4-methylphenyl)methanol	✓ methoxybenzene	✓ methoxybenzene	1. Structure, links & more	2. Abstract, te & more

End of this inhibitor

1G2K
Synthesis and comparative molecular field analysis (CoMFA) of symmetric and nonsymmetric cyclic sulfamide HIV-1 protease inhibitors.

Schaal W, Karlsson A, Ahlsen G, Lindberg J, Andersson HO, Danielson UH, Classon B, Unge T, Samuelsson B, Hulten J, Hallberg A, Karlen A.

J Med Chem. 2001 Jan 18;44(2):155-69. Abstract
Related 2-D structures, Cellular data, Anti-HIV enzyme data, References

Restart from 1G2K

Fragments of 1G2K. (Click another fragment to search inhibitors using an additional fragment)

1,2,7-thiadiazepane-4,5-diol 1,1-dioxide	3-methylbenzaldehyde	PHENYL	✓ methoxybenzene	✓ methoxybenzene	methylamine	1. Structure, links & more	2. Abstract, te & more

Fig. 5. The result of query with search engine type II for compounds that contain methoxy benzene. This query is done by the hyperlink on methoxybenze in Figure 4. For brevity, only a part of the result page is shown. The fragments that were already used in the substructure search are shown with a tick mark, and they are not hyperlinked.

Search engines type I

The main purpose of this search engine is to query compounds by using a substructure (fragments or groups or classes) of a compound (Fig. 1, "distant relative" or "more distant relative") defined in layer 1, 2, or 3, respectively. This search engine displays data from a given layer as a hyperlink. When one of these hyperlinks is chosen, it displays data from the next layer (column of a data table) that corresponds to that particular data. Layer 1 of the data tree is populated with the class of a (fused rings, sulfur-containing groups, six-member rings, etc.) group stored in layer 2 (column 2) of the same row of the database table. At layer 1, this search engine allows a user to choose one of these classes; for instance, a six-member ring that produces a list of all the groups that contain six-member rings (e.g., phenyl, Fig. 3). At layer 2, it allows one to make a selection on a group, and at layer 3, it allows one to make a selection on a fragment. Thus, the hyperlink on the fragment, methoxybenzene (Fig. 4.) gets all the compounds with this fragment (Fig. 5), and at the end of layer 3, control is passed on to search engine type II to facilitate query on multiple fragments. A novel feature of this search engine is its ability to present the data to users in several layers. In each layer, it facilitates the definition of a probe to specify the target compound. There are >1000 chemical fragments in our database. These search engines present only a few of these fragments at a time. Using these layers, a user is able to filter out most of the unwanted compounds in a couple of steps.

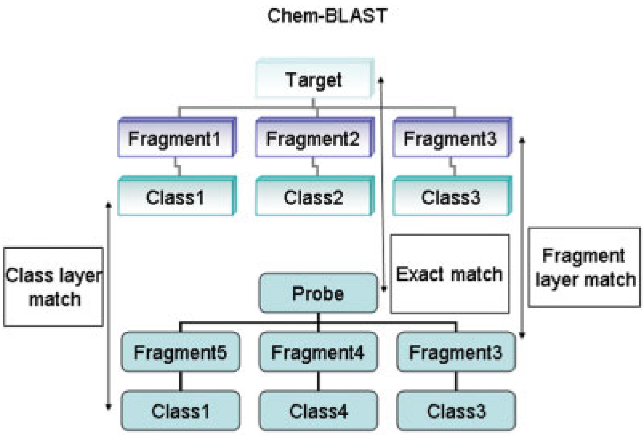


Fig. 6. The principles of Chem-BLAST for searching and aligning chemical structures. For brevity, only two layers are shown. The method queries on related compounds by matching data from corresponding layers of target and probe compounds. When all fragments of a probe and target match, identical compounds result. When only fewer fragments match, related compounds result. When only classes match, distantly related compounds result.

Search engine type II

This search engine defines and initiates search for related compounds that have several structurally independent chemical fragments in common (Fig. 1, "close relative"). It presents data from a lower layer as hyperlinks in a Web page that is displaying data from higher layers. These higher layer data are provided to inspect compounds, and hyperlinks on lower layer data are provided to select one of its substructures to define its structural neighbor.

Comparison with other search engines

Many search engines (e.g., as in NIST Chemistry WebBook[3]) are available where users can produce hit lists based on ad hoc substructures and refine these hits further by using modified substructures. Some of the benefits of a data tree-based database over such databases are summarized in Table I. For a moment, consider a biological database that can be provided either in a free text-based ad hoc query string-driven version (PDB) or in a "taxonomy"-based data tree-driven version (e.g., http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/). Certainly, it takes more effort to annotate and create a taxonomy version than a free text-based version of a database. However, taxonomy-based versions are increasingly popular and novel. These taxonomy databases provide results in an orderly fashion with predictable, complete, and manageable results using concepts commonly used by researchers. The data tree approach proposed here for compounds is such a taxonomy version of a structural database.

Medicinal chemists often group compounds by using their fragments like biochemists who group and align proteins by using their amino acid sequences with the program BLAST.[20,21] Here we describe a novel method, Chem-BLAST, standing for Chemical Block Layered Alignment of Substructure Technique (Fig. 6), to look for structural neighbors with similar fragments (fragment, subgroup, or class). Search engines use this Chem-BLAST to support query on compounds using one or more (multiple probes define a motif) chemical fragments. In this technique, a chemical motif may also be defined by choosing multiple fragments either in successive queries or by concatenating names of fragments using "AND" or "OR" operators (Fig. 7). For convenience, these search engines allow users to specify fragments for a query either by using hyperlinked molecular drawings or using IUPAC names.

Query of Related Structures

Query on related structures may be defined as a query that produces a list of structures that have one or more substructures in common. As explained earlier, a query performed in "zoomed out" state does exactly that. In our method, substructures are defined and indexed in layers (fragments, groups, and class) of the data tree. The substructures in these three layers provide three distinct (close to distant, Fig. 1) ways to define structural relationships. Layers 1 and 2 provide general structural features, whereas layer 3 provides greater structural details to define structural neighbors.

Searching neighbors with a substructure

With the use of a data tree, a task of searching neighbors with a particular substructure amounts to displaying all the structures in layer 4 of the data tree, which have a given substructure in a lower layer. Imagine that one wants to use six-member rings (Fig. 1, search on substructure) to search a set of compounds. For this purpose, by using a data tree, the search engine displays all six-member rings (layer 2, Figs. 1 and 3). By using any one of these hyperlinks, a user may query the database and obtain a list of all the fragments that have six-member rings (layer 3, Figs. 1 and 4). From this list, he may choose a fragment of his interest and get a list of all the compounds (layer 4) with that fragment. Additional fragments may be appended to the query at this stage by using search engine type II. In the absence of a data tree, a user needs to guess the types of six-member rings that may exist in the database and test them one by one through queries on each of them. Thus, in the absence of a data tree, each of these steps would have been more complicated because of the difficulty in obtaining the information on what is available in the database. Search engines, such as in Relibase or the PDB, do not have a direct mechanism to let the user know what substructures are available in the database for substructure searches. Such search engines[2] start off by asking a user what he wants rather than telling a user what is there in the database to search on. Such search engines are like "pharmacies" that serve prescriptions rather than like "supermarkets" where things are laid out for you to pick and choose.

Searching neighbors starting from a compound

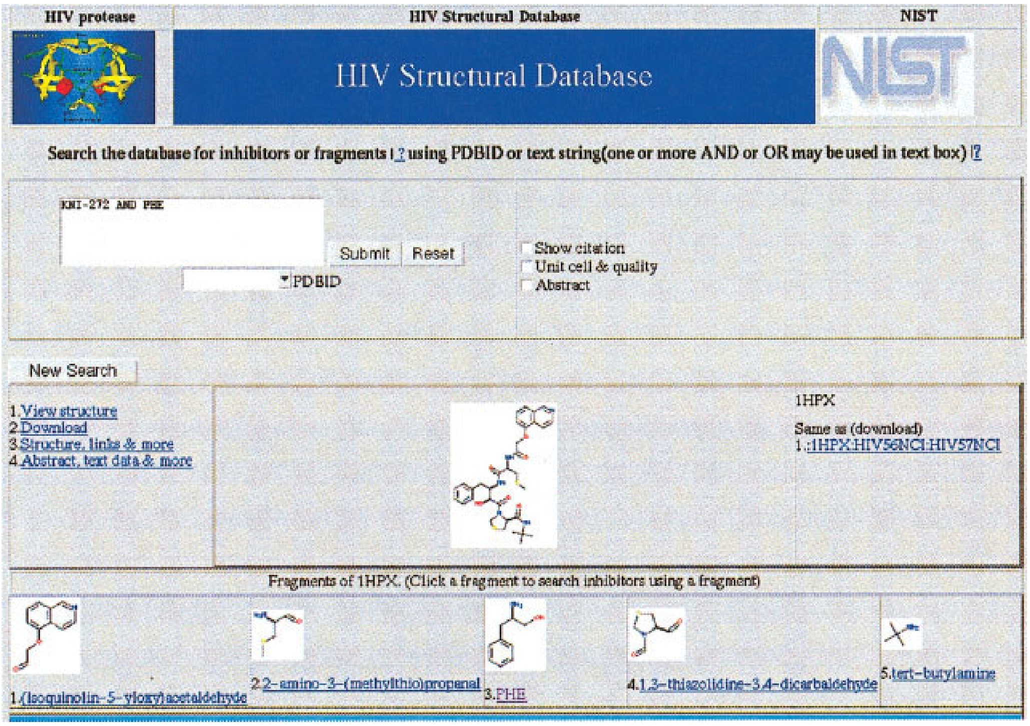
The method described above for searching neighbor using a substructure provides an indirect way of obtaining related structures. A more direct way of obtaining structural neighbors is based on the method used by a molecular biologist who begins with a protein and looks for related proteins by using its certain predefined characteristic amino acid sequences. In this method, one starts from a compound and asks the question: give me all the structures that are related to it as defined by its unique features. An illustration of such a method for compounds is available in the PubChem (http://pubchem.ncbi.nlm.nih.gov/search/); this illustration lists structural neighbors by using criteria preestablished by the database providers. However, structural features of interest to a user often cannot be defined in advance for a compound and may vary depending on the user and his interest at a given time. To take care of this problem, one may rephrase the question on structural neighbors as the following: give me all the structures related to a structure by substructure(s) of my choosing. Enabling such a query is difficult in the method used by the PubChem where structural neighbors are defined in advance. However, in the presence of a data tree, search engines described in this work are able to support such a query as well, by presenting the user with

TABLE I. Comparison of Chem-BLAST, Chemical Taxonomy Method With Other Methods

Task	Chem-BLAST	Other methods
Data preparation and Web development	Develop special tools and data dictionaries to annotate compounds into a data tree.	Develop special tools to support input of 2D structures of a compound on a Web.
	Develop tools using standard database software, such as Oracle or MySQL, to query both on the elements of a data tree and other data on compounds.	Develop tools using SMILES to query on 2D structures. Develop tools to query other data using data bases, such as Oracle or MySQL.
	Incorporate expert value added information during the annotation for individual substructures of each compound.	
	Develop intelligent "semantic" Web tools using the expert value added information on the sub-structures of each compound.	
	Efficient use of data resource because both data storage and query are done by using commercial software, such as Oracle, used in many other disciplines.	Needs special software to enable drawing and interpreting 2D structures as input and query.
	Efficient query using substructures that are indexed in advance by using data tree.	Database may not be preindexed on substructures as they are generated and searched on the fly for each usage.
	Indexing of substructures is done only once during the data annotation; thus, better use of computational resource is achieved with more reliable hits.	Interpretation and indexing of substructures is done individually for every query. Therefore, one needs better Web tools and more computational power to reduce missed hits.
	Intelligent "data-aware" tools, similar to taxonomy-based tools often used for biological text data, may be developed by using the layered indices of the data tree.	Only limited "data-aware" tools may be developed because of lack of layered indices on substructures.
	Excessive and missed hits may be controlled. User may be given a "guided tour" of the substructures by using layers of the data tree.	Too general substructure for a query may produce excessive hits, and too specific substructure for a query may re sult in missed hits.
Substructure query	Structural features available in the database for query are laid out in 2D to pick and choose like groceries in a supermarket.	Structural features for a query are not laid out for a user to examine; a user postulates and draws a substructure for query. The situation is like that of a user describing his needs to a store keeper who picks items for the user based on his best guess.
	By using one or two mouse clicks, a user may obtain list of all the substructures available in the database.	A user may not know what substructures are available because the available ones are not laid out to examine.
	Queries may be composed quickly by using hyperlinks of the substructures.	A user needs to draw the substructure dynamically, which may need tens of mouse clicks and considerably more expertise.
Structural neighbors	Structural neighbors of a compound may be quickly queried by using hyperlinks on its substructures.	Query on structural neighbors is labored because it is not synchronized with the display of a structure. A user is expected first to examine a structure and then draw its substructure for search on its structural neighbors.
Structural motif	Multiple substructures may be used to specify a motif among structural neighbors.	

options for choosing substructure(s) in one or more steps. In our implementation of the data tree, inhibitors may be searched either by using a substructure as described above (Fig. 5) or by using a text input (Fig. 7). In either of these cases, substructures for defining structural neighbors may be chosen through any of the hyperlinks on structural fragments displayed underneath each one of these structures. These hyperlinks produce a list of compounds that have a given fragment in common. Additional fragments to define structural neighbors may be chosen in a subsequent Web page. The hyperlinks disappear when query has converged to produce only a few hits.

The basic advantage of the proposed method stems from the fact that the total number of substructures at the lower layers of a data tree is smaller than at higher layers; therefore, the search engines are able to produce manageable hits for a query for substructures in lower layers. The prescreening of the hits at lower layers by a user results in



HIV protease | HIV Structural Database | NIST

HIV Structural Database

Search the database for inhibitors or fragments ! ? using PDBID or text string(one or more AND or OR may be used in text box) !?

KNI-272 AND PHE

Submit Reset

PDBID

□ Show citation
□ Unit cell & quality
□ Abstract

New Search

1.View structure
2.Download
3.Structure, links & more
4.Abstract, text data & more

1HPX

Same as (download)
1.:1HPX:HIV56NCI:HIV57NCI

Fragments of 1HPX. (Click a fragment to search inhibitors using a fragment)

1.(isoquinolin-5-yloxy)acetaldehyde
2.2-amino-3-(methylthio)propanal
3.PHE
4.1,3-thiazolidine-3,4-dicarbaldehyde
5.tert-butylamine

Fig. 7. The use of "AND" in a query. This feature may be used to specify multiple fragments to specify a structural motif within a probe. This Web page was generated with search engine type II, and it shows a compound and its five fragments as hyperlinks for additional query. A user may use any one of the five hyperlinks to get a list of all the compounds with that fragment in common between them. The structural homology among the hit list may be improved by specifying additional fragments in a subsequent query.

manageable hits even at higher layers. Furthermore, all the substructures for query are predefined in the data tree; therefore, substructures could be indexed in advance to produce hits in an orderly and predictable fashion. Moreover, by using the data tree, search engines are able to produce hits both in reverse (structure to substructure) and forward (substructure to structure) direction to facilitate explicit homology searches of a structure using structural features chosen by a user. In the absence of a data tree, search engines propagate queries on inputs that are based on users' experience and intuition; thus, homology searches are much more ad hoc and tedious. Novel predictability and reversibility of a structure-based query are used by search engines to produce a list of related structures using the criteria dynamically defined by a user through their substructures. In our application, whenever possible, fragments were generated by breaking at peptide-like bonds; thus, many of the fragments are "semantically defined" for HIV protease. For other applications, semantically defined substructures may be developed by using an appropriate data dictionary to define substructures.

DISCUSSION AND CONCLUSIONS

We are witnessing the emergence of a Web-based "data-rich" era for chemical and biological compounds. In the past decade, databases have become an integral part of research and development in the biomedical sciences. Bioinformatics now plays an essential role both in deciphering genomic, transcriptomic, and proteomic data generated by high-throughput experimental technologies, and in organizing information gathered from traditional biology. To make significant advances in this data-rich era, it is essential to introduce techniques that allow interoperable annotation, query, and analysis across diverse data; plug-and-play scalable annotation, and adoptive query tools that facilitate seamless interplay of tools and data; and versatile user interfaces that allow researchers to annotate, visualize, and present the results of analysis in the most intuitive and user-friendly manner.

Scalable Annotation

Right now there is not a single international standard for naming compounds and their fragments uniformly,[10] and the current standardization efforts using InChI[5] are focused on entire compounds.[5] However, most query and comparison tools (e.g., WebBook, Relibase http://relibase.ebi.ac.uk/reli-cgi/rll?/reli-cgi/general_layout.pl?home) for compounds are based on certain fragments of a compound. Thus, a mismatch exists between the internal representation of compounds in a database and in their usage to query structures on a Web. The method described here decomposes compounds into standard fragments by using context-dependent data dictionaries. Such partitioning of a large chemical into smaller fragments may not always be unambiguous. For this reason, a multipronged approach using data dictionaries and semantic Web technologies is suggested. In this approach, compounds are grouped into multiple fragments by using several alternative data dictionaries, which include synonyms of structural fragments. These definitions are then used by automated procedures to perform annotation and define data tree and semantics to support query for structural neighbors.

Adoptive Query Tools

Our annotation method and Web tools provide a number of novel features both for annotation, query, and display of chemical structure. During annotation, this method establishes data definitions in several layers of chemical information on fragments. These layers are used by search engines for presenting the data in an intuitive manner. A Web user makes decisions in several steps, and in each step, he refines his query to define the results with increased accuracy. This layered approach to refine a probe reduces the need for prior in-depth knowledge of the fragments of compounds that are available in the database. The method allows a user to learn about available options in each layer and to define or refine his probe accordingly. The method is related to the "heap sort method," used for sorting data. This technique reduces the number of comparisons from N^2 to $N \log_2 N$ steps for a binary tree. Large databases that hold and distribute complex structures that may be fragmented into smaller pieces for query are the primary focus of our method. HIV protease-inhibitors[5,14] have about five fragments per inhibitor (about 40 non-hydrogen atoms). By using the proposed Chem-BLAST, usually a compound may be specified and located in this database within a couple of "mouse clicks." The proposed annotation enables the attachment of "meaning" (i.e., semantics) to compounds in a manner that far exceeds the current practice of associating "metadata" with data for compounds. This is accomplished by creating a knowledge base (or ontology) associated with compounds. One defines "concepts" in terms of primitives in such a way that taxonomies can be inferred, thus, significantly reducing the size of the database to enrich annotation and query capabilities for drug design purposes.

The popular paradigm in drug discovery seeks to collect, compare, and test many chemically similar compounds. This approach of modern drug discovery[22] is rational and knowledge based with a defined hypothesis on the functional role of individual fragments that make up a drug. The process of drug design, therefore, begins with a lead compound followed by a hypothesis about how different fragments of this compound interact with the amino acid residues of the target protein molecule.[23] Following this, database searches are performed to gather structural neighbors of the lead compounds by using what is commonly known as "mix and match method of the functional fragments" of a lead compound. The Chem-BLAST, proposed here for this purpose, is similar to the commonly used BLAST[21] to search for sequence neighbors. Several databases are available to search through chemical data, but the ability to search for structural neighbors of particular compound held in these databases is far from satisfactory. Some of these databases[2,3] allow searches using arbitrarily assigned index or IUPAC names that are hard to memorize. Some Web sites (e.g., Relibase) facilitate drawing the probe interactively. However, this method has the following limitations:

1) Drawing a complex probe on the fly is difficult.
2) The probes are often drawn by a user in the absence of tools to obtain accurate knowledge on what is available in the database. Too restrictive probes may miss hits, and too general probes produce overwhelming hits. Any mismatch between the probe and the probed, for instance C=O instead of C—O produces unexpected result. Such Web tools are "semantically blind," and they do not attach or enforce implicit "meaning" to a probe for the target compounds held in the database.
3) Some databases, in an attempt to overcome this problem, provide chemical templates to aid the choice of a probe. But these templates may not be "semantically aware": the database need not necessarily have compounds that match a probe built from these templates. These templates do not have a direct relationship to the compounds held in the database.
4) The data annotation procedures adopted by many of these databases[2,24] do not establish the structural boundaries among fragments to allow search on a target using multiple probes, such as Phe and thioproline of KNI-272, which are resolved in space.

The method proposed here tries to overcome most of these limitations. It establishes standard fragments that have a direct mapping to the compounds and thus enables the implicit attachment of metadata to an object that represents the compounds held in the database. By using these fragments, the Chem-BLAST performs single or nonoverlapping multiprobe search on fragments to list structural neighbors. In summary, here we present a method that lays the foundation for the use of Semantic Web Technology on compounds; it allows the definition of probes that have a structural "meaning" (Figs. 1 and 2) for the compounds held in the database; it guides a user in the selection of probes that are guaranteed to produce hits in the database; it also facilitates the choice of multiple probes that may be used to reduce unwanted hits by defining targets more accurately. In our method, structures are implicitly connected among each other in the database through the standard fragments, thus enabling the use of predefined fragments as probes to search for structural neighbors.

ACKNOWLEDGMENTS

We thank F. Schwarz, R. Goldberg, S. Stein, G. Gilliland, M. Tung, and V. Vilker for helpful discussions and support. Partial funding of this effort was provided by the Systems Integration for Manufacturing Applications (SIMA) and also by the Exploratory Research Award for the year 2004, and by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

REFERENCES

1. Bhat TN, Bourne PE, Feng Z, Gilliland GL, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H, Westbrook J,

Berman HM. The PDB data uniformity project. Nucleic Acids Res 2001;29:214–218.

2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
3. Linstrom P, Mallard WG, editors. NIST Chemistry WebBook, NIST Standard Reference Database. Gaithersburg, MD: National Institute of Standards and Technology; 2003.
4. Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm W, Weissig H, Greer DS, Bourne PE, Berman HM. The Protein Data Bank: unifying the archive. Nucleic Acids Res 2002;30:245–248.
5. Prasanna M, Vondrasek J, Wlodawer A, Bhat TN. Application of InChI to curate, index and query 3-D structures. Proteins 2005;60:1–4.
6. Joseph-McCarthy D. Computational approaches to structure-based ligand design. Pharmacol Ther 1999;84:179–191.
7. Wlodawer A, Erickson JW. Structure-based inhibitors of HIV-1 protease. Annu Rev Biochem 1993;62:543–585.
8. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL. Nucleic Acids Res 2003;31:365–370.
9. http://www.acdlabs.com/products/chem_dsn_lab/chemsketch/tech.html ACD/ChemSketch IUPAC names.
10. Adoms D. Chemists synthesize a single naming system. Nature 2002;417:369.
11. Bemis GW, Murcko MA. The properties of known drugs. I. Molecular frameworks. J Med Chem 1996;39:2887–2893.
12. Whittle PJ, Blundell TL. Protein structure-based drug design. Annu Rev Biophys Biomol Struct 1994;23:349–379.
13. Feldmann RJ, Milne GWA, Heller SR, Fein A, Miller JA, Koch B. An interactive substructure search system. J Chem Inf Comput Sci 1977;17:157–163.
14. Vondrasek J, Wlodawer A. HIVdb: a database of the structures of human immunodeficiency virus protease. Proteins 2002;49:429–431.
15. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. J Chem Inf Comput Sci 1998;38:511–522.
16. Adamson GW, Cowell J, Lynch MF, Mclure AHW, Town WG, Yapp AM. Strategic consideration in the design of a screening system for substructure searches of chemical structure files. J Chem Document 1973;13:153 – 157.
17. Bhat TN, Baldwin ET, Liu B, Cheng YS, Erickson JW. X-ray structure of a tethered dimer for HIV-1 protease. Adv Exp Med Biol 1995;362:439–444.
18. Baldwin ET, Bhat TN, Gulnik S, Liu B, Topol IA, Kiso Y, Mimoto T, Mitsuya H, Erickson JW. Structure of HIV-1 protease with KNI-272, a tight-binding transition-state analog containing allophenylnorstatine. Structure 1995;3:581–590.
19. Stein SE, Tchekhovskoi D, Heller SR. 2004 http://www.iupac.org/projects/2000/2000-025-1-800.html
20. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003;31:23–27.
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
22. Drews J. Drug discovery: a historical perspective. Science 2000;287: 1960–1964.
23. Kuntz I. Structure-based strategies for drug design and discovery. Science 1992;257:1078 – 1082.
24. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland GL, Weissig H, Westbrook J. The PDB and the challenge of structural genomics. Nat Struct Biol 2000;11:957–959.