On the Creation of Structural FaceBook using Rule-Based Methods to Build and Exchange Ontology for Drug Design

Talapady. N. Bhat
Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA.
bhat@nist.gov

Abstract: The AIDS HIV structural databases (HIVSDB, http://bioinfo.nist.gov/SemanticWeb_pr2d/chemblast.do), the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) and the PubChem (http://pubchem.ncbi.nlm.nih.gov/) distribute one of the largest comprehensive collections of structural data on inhibitors, drug leads and clinical drugs for many diseases including AIDS. These databases contain info on several thousand biologically active compounds from many classes (HIV PR, RT, CCR5, Integrase) of FDA approved drugs for AIDS. Efficient and yet user friendly rule-based data management systems that support state-of-the-art annotation, visualization and query capabilities are crucial for the effective use of data for fragment based structural pharmacology and rational drug design. Semantic Web is the vision of the World Wide Web Consortium for enabling seamless integration of electronic data for data mining and knowledge generation across the Web. Robust and functionally relevant ontology plays a critical role in developing the data elements for a Semantic Web. Presentation will illustrate how Rule-based Semantic Web concepts are used for novel annotation, data integration, storage, and query to manage and display structural (fragments, 2-D images and text-based) biological, and pre-clinical data (http://xpdb.nist.gov/chemblast/pdb.html). The technique is called (Chem-BLAST – Chemical Block Layered Alignment of Substructure Technique) and it allows rapid comparison and exchange of compound information using automated rule-based methods to develop structural ontology for use in drug discovery process. The methods and the results that will be presented are probably the first of its kind in biological world that uses entirely rule-based event processing methods to build, integrate and exchange information. The ontology developed by the methods is amenable to be presented either as an RDF or XML or to be used in a relational database. We call it a structural facebook as it is driven by ontology of structures of 'shared features' and presents them using visual images for rapid comparison.

Key-Words: Bioinformatics: the future; Knowledge-based applications in structural Chemistry for AIDS; Structure-Based Drug Design, the Structural Informatics

1 Introduction

Recent developments in structure-based drug-design, combinatorial chemistry, structure genomics and high-throughput screening have resulted in several large databases such as the Protein Data Bank (PDB)[1], HIV structural Database (HIVSDB) [2], SwisProt/ChEBI [3] and the PubChem [4] with data on ligands, drug-like molecules and their complexes with drug targets. The number of compounds held in these

databases range from tens of thousands to millions. Many compounds held in these databases are structurally similar. Further, many compounds are also common among these databases.

Efficient use and seamless integration of data held in such large databases is a daily problem for their users. This situation has created an urgent need for robust rule-based techniques for pharmacological and structural profiling of large number of compounds within single or across multiple databases. Further, chemical compounds, both because of their over-whelming number and technological importance are one of the ideal candidates to develop, test and illustrate the novelty of rule-based techniques in bringing 'order into a chaotic world'. In this paper we describe and illustrate how such a technique can be designed, developed and then utilized to build and exchange knowledge for the compounds held in the PDB, HIVSDB and the PubChem.

2 Preamble for the Rules

Structural biologists and medicinal chemists tend to consider a chemical compound as a collection of its components[5-8] and these components are commonly known as substructures. This consideration is analogous to the consideration of a machine by an engineer as made up of several independently definable components. A medicinal chemist may use this concept to iteratively modify a compound in certain locations to manipulate certain properties. A molecular modeler may generate new model compounds by iteratively substituting substructures. A structural biologist may elucidate enzyme drug interaction, for instance for studying drug resistance, by examining enzyme-drug complexes of compounds of similar substructures. The method described here provides a rule-based approach to define and generate substructures and then build substructure based ontology for compounds. This ontology is organized and indexed in a way that allows its efficient use over a Web that may also be used in a Semantic Web. Searches on thousands or millions of compounds can also be done rapidly using this ontology. The ontology also allows the use of ontological terms to perform accurate or fuzzy searches. Accurate compound similarity may be defined by terms that ensure that all atoms and their bonds be identical. Fuzzy similarity may be defined by ensuring that only a few of the many substructures are preserved among the resulting compounds. This method builds a ontology of substructures using rules that operate on chemical blocks of layers of aligned template of substructures of compounds of interest. The method that we describe here is called Chemical Block Layered Alignment of Substructure Technique (Chem-BLAST).

2.1 Rules

The proposed method generates ontology in several successive steps by applying rules that operate on the atomic connectivity of a compound. For the sake of simplicity, we describe here only few of the rules and the steps that implement them in Chem-BLAST.

2.2.1 Step 1

The first step of Chem-BLAST generates level 1 of the ontology by performing a mapping of the substructures of a target compound to a set of rules that define a substructure. For, simplicity here we chose to limit our discussion on the rules that are applicable only to two types of rings – a six member ring and a double fused ring. A set of atoms are defined to belong to a substructure called six member ring if (a) every atom in that set is connected to at least two other atoms of that set; (b) starting from any atom of that set, on stepping through successively six times over connected atoms, one gets back to the starting atom of that set; (c) the total number of atoms in that set is six. If all the atoms in this ring are carbon atoms, it is called a carbon-containing six member ring (Fig. 1).

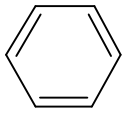

Fig. 1 shows a six-member ring made of only carbon atoms. Each corner of this hexagon shaped structure shows the location of a carbon atom and the line (both single and double) between these corners show the type of atomic connectivity between these atoms. In this ring, every atom is connected to two other atoms and starting from any atom of this ring, one can step through six bonds to get back to the starting atom.

If the number of atoms one need to step though in a ring is always a constant and it is equal to n, then it is called n-member ring (where n is the number of atoms in the set), other-wise it is called a fused ring (Fig 2). A fused ring is also expected to have one or more bonds that are shared between two or more adjacent rings.

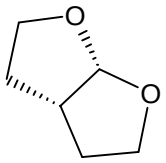

Fig. 2 shows an oxygen-containing double-fused ring made up of two five-member rings.

2.1.2 Step 2

Following the mapping of a target compound to a level, called level1 (step 1) in this paper, compounds are mapped into level2 (step 2). Atoms of level2 consist of all the atoms of level1 plus their neighboring atoms defined as follows. The neighboring atoms are defined as those atoms that are bonded to an atom of level1 but it is not already a part of level1 (Fig. 3).

2.1.3 Step 3

Mapping of compounds into new levels with rules similar to that of step 2 are continued until all the atoms of the compound are included in the final set (Fig. 3).

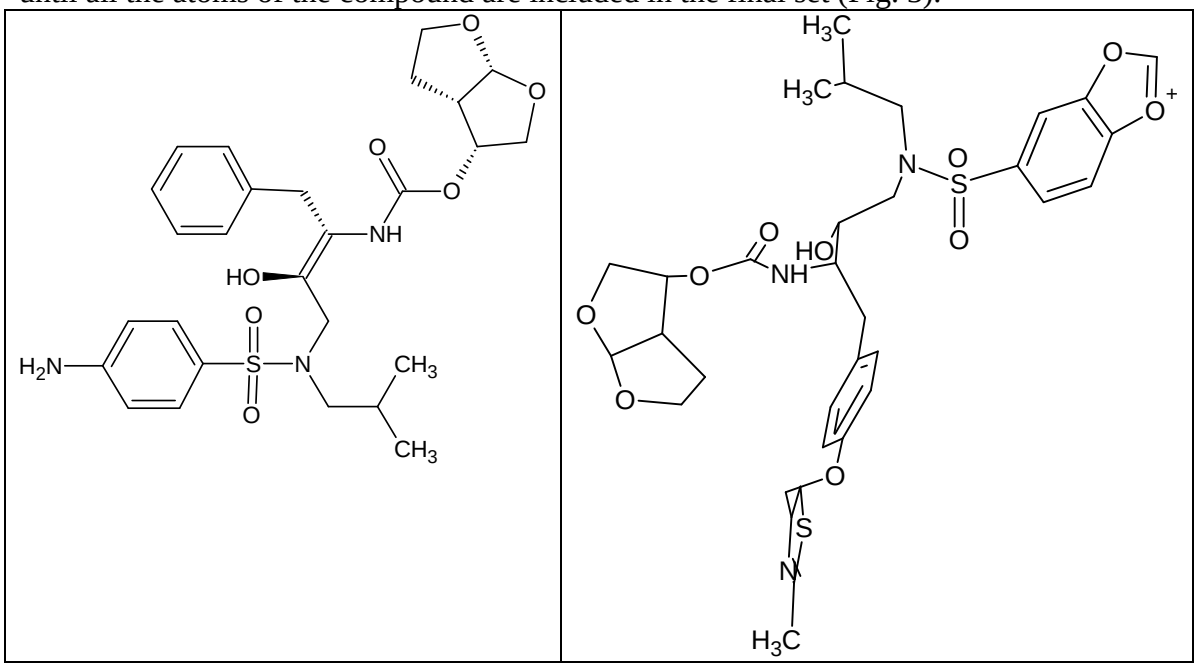

Fig. 3 shows two compounds that contain the substructures shown in Fig. 1 & 2. The compound to the left is know as TMC114 which is an inhibitor of HIV Protease (PDBID 1T3R). HIV protease is a target for AIDS drug design and about half of the drugs currently used to combat AIDS belong to this class of drugs. Further details on this compound are available at http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=1T3R . The compound to the right is also an inhibitor of HIV protease (PDB ID 2FDD).

2.1.4 Step 4

After identifying the substructures (Steps 1- 3), the method maps all the substructures and the compounds into a set of target templates of unique substructures and compounds. In the above example (Fig. 1-3), such unique template is made of two substructures and two compounds. This information is then expressed as a set of graphs (Fig 4). This set of graphs collectively establishes an ontology that contains information on all the compounds and their substructures. Additional information such as the biological data and the significance of the compound with respect to use-cases such as drug-discovery is also added to the set of graphs. All the graphs collectively function as the ontology.

3 Implementation/discussion

To test and illustrate the concept of Chem_BLAST, we have developed ontology and a Web based resource for information on about a million compounds chosen from the PDB, HIVSDB and PubChem (Fig. 6). This resource operates by first a user specifies a substructure for query chosen from one of the elements on the graph. Names of substructures or of compounds are notoriously famous in their lack of clarity and uniformity and for this reason alone name-based queries of chemical structures have been frustrating for user. To overcome this problem, Chem-BLAST displays the elements of the graphs in a series of layers (layer1 to layer n) of molecular images of the substructure they denote. This technique of using predefined molecular images of the substructures to choose from for a query overcomes the difficulties of establishing text-based standard intuitive names for the elements used in the ontology. The element specified by the user is then compared to all the target elements in that level (hub) and the elements from the next level (spoke) are displayed as image for further selection by the user. All the hubs and the spokes that are available in the ontology are pre-indexed and held in a database such as ORACLE or MySQL and thus the response to a user query can be rapid and precise.

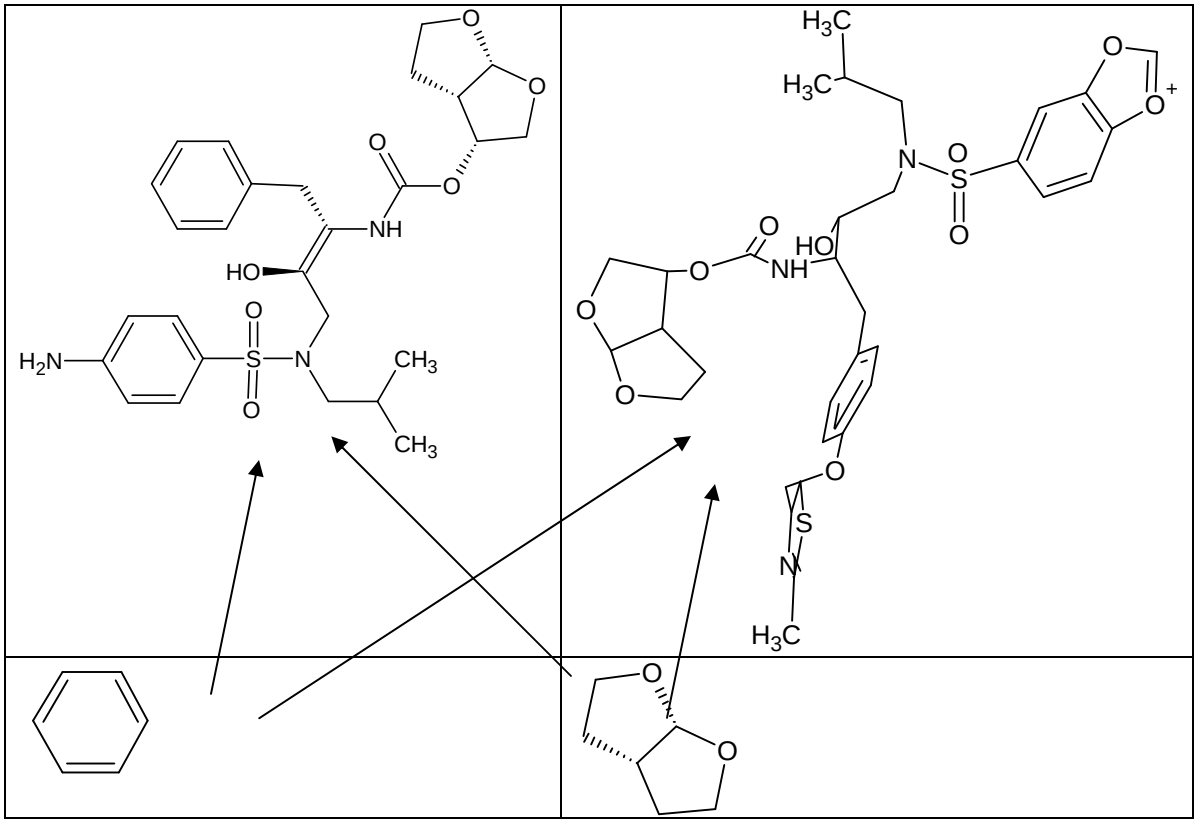

Fig. 4 shows a graph that denotes relationship between two substructures (bottom row) and two compounds (top row).

3.1 Chem-BLAST and RDF

Chem-BLAST annotates data into a set of graphs (Fig. 4) of uniformly named "begin and end' points. The graphs are expressible as RDF of the type "begin" – "is a part of" - "end". Since each structural element denoted by a 'begin' or 'end' is obtained by mapping substructures to a template, it is unique too. For this reason, the graphs may also be loaded and queried using SPARQL based databases and exchanged by web services that require unique identifiers for each data element.

3.2 Chem-BLAST and Relational Database

Relational databases are the preferred choices for managing data; particularly for large databases such those of chemical compounds discussed here. For this reason, Chem-BLAST re-organizes the graphs into data-trees, and we illustrate this process of re-organization of data for a limited set of rings (Fig. 5). The topmost hub of the tree shown in Fig. 5 is of "Rings". This hub has two spokes called "Simple rings" and the other is a "Fused rings". The spoke "Simple rings" has one branch – a ring with six carbon atoms. There are two compounds in this fictitious database and each compound has one or more rings with six carbon atoms. Each hub and spoke of this tree is identified uniquely and therefore it may be indexed using the standard database vender provided techniques to support efficient queries.

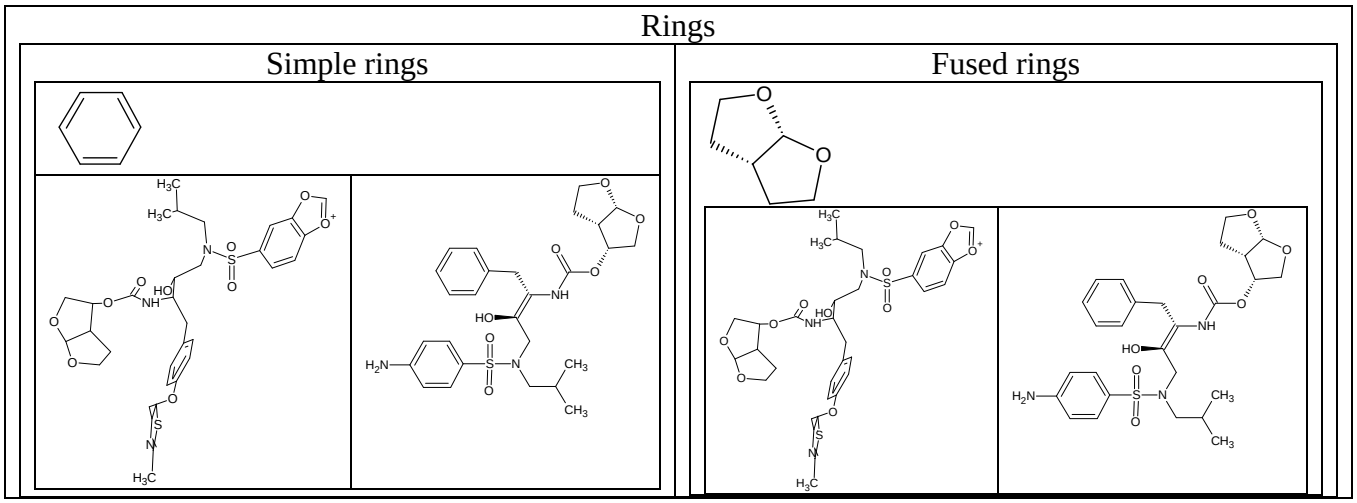

Fig. 5 shows a tree of data generated by Chem-BLAST for two compounds. The topmost hub of the tree is "Rings" and it branches out to two types of rings – Simple rings and Fused rings, and then to a particular ring and then to two compounds.

3.3 Chem-BLAST and Semantic Web

The vision of Chem-BLAST is to help to enable seamless exchange and integration of knowledge on chemical compounds over the Web. For this purpose, Chem-BLAST uniquely identifies each element of the RDF (3.1) by mapping it to a standard template. These identifiers are an encoded form of the atomic connectivity of the atoms[2, 9] of the substructure. This encoding uses a standard rule accepted by the International Union of Pure and Applied Chemistry (IUPAC). Therefore, in principle, researchers may be able to use this technology to prepare/exchange data using chemical Semantic Web concepts.

.

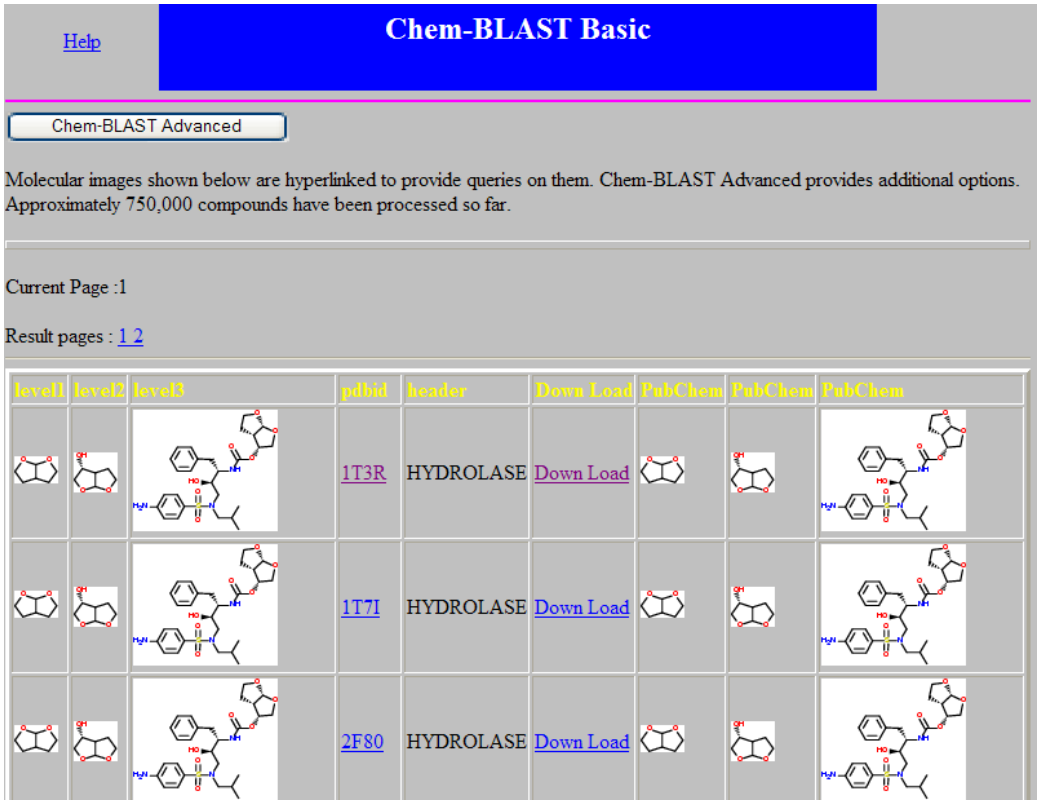

Fig. 5 shows (http://xpdb.nist.gov/chemblast/pdblevel1.pl?T1=1T3R_l1_2) a part of the Web interface with the results of a query on the double-fused ring discussed above (Fig. 2). Using this page a user may intersect structures from the PDB (left) and PubChem (right) by clicking the hyperlinks associated with the molecular images of the substructures (columns 1,2 for the PDB and 7, 8 for the PubChem) or the compound (columns 3 or 9) that he/she wants to use to intersect between the databases. For instance, if the user clicks the double fused ring under PubChem, all compounds (at this time about 750,000 compounds from PubChem have been included in the database) with that substructure will be displayed from PubChem, or if he/she clicks the double fused ring

from the PDB, he/she gets all the compounds with that substructure from the PDB. The Webpage created by Chem-BLAST is called structural facebook as it is built around a structural ontology that groups and facilitates intersection of structures using familiar concepts as described above.

Disclaimer

Certain trade and company products are identified in this paper 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), or does it imply that the products are necessarily the best available for the purpose.

References

1. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
2. Prasanna, M., Vondrasek, J., Wlodawer, A., Bhat, TN., *Application of InChI to curate, index and query 3-D structures.* Proteins, Structure, Function, and Bioinformatics, 2005. **60**: p. 1-4.
3. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL.* Nucleic Acids Res., 2003. **31**: p. 365-370.
4. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2007. **35**(Database issue): p. D5-12.
5. Lewell, X.Q., et al., *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**(3): p. 511-22.
6. Debnath, A.K., *Application of 3D-QSAR techniques in anti-HIV-1 drug design--an overview.* Curr Pharm Des, 2005. **11**(24): p. 3091-110.
7. Rummey, C., et al., *In silico fragment-based discovery of DPP-IV S1 pocket binders.* Bioorg Med Chem Lett, 2006. **16**(5): p. 1405-9.
8. Prasanna, M.D., et al., *Chemical compound navigator: a web-based chem-BLAST, chemical taxonomy-based search engine for browsing compounds.* Proteins, 2006. **63**(4): p. 907-17.
9. Murray-Rust, P., et al., *A global resource for computational chemistry.* J Mol Model (Online), 2005. **11**(6): p. 532-41.