Building Chemical Ontology for Semantic Web Using Substructures Created by Chem-BLAST

Talapady N. Bhat, National Institute of Standards and Technology, USA

ABSTRACT

Efficient and user friendly ontologies are crucial for the effective use of chemical structural data on compounds. This paper describes an automated technique to create a structural ontology for compounds like ligands, co-factors and inhibitors of protein and DNA molecules using a technique developed from Perl scripts, which use a relational database for input and output, called Chem-BLAST (Chemical Block Layered Alignment of Substructure Technique). This technique recursively identifies substructures using rules that operate on the atomic connectivity of compounds. Substructures obtained from the compounds are compared to generate a data model expressed as triples. A chemical ontology of the substructures is made up of numerous interconnected 'hubs-and-spokes' is generated in the form of a data tree. This data-tree is used in a Web interface to allow users to zoom into compounds of interest by stepping through the hubs from the top to the bottom of the data-tree. The technique has been applied for (a) 2-D and 3-D structural data for AIDS[1]; (b) ~60,000 structures from the PDB [2,3]. Recently, this technique has been applied to approximately 3,000,000 compounds from PubChem[4],[5,6]. Plausible ways to use this data model for the Semantic Web are also discussed.

Keywords: *Chem-BLAST, Chemi-Informatics, Chemical Ontology, Ontological URI (OURI), Structural Informatics, Structure-Based Drug Design, URI*

1. INTRODUCTION

Established methods, e.g. PSI-BLAST (Altschul et al., 1997) that use amino-acid sequences to compare and organize structural data are widely used to build structural ontology for protein and DNA molecules. These proteins and DNA molecules are often found to contain small molecules such as toxins, drugs, and co-factors. These small molecules fall under a broad category called ligands (*compounds*) in the Protein Data Bank. *Compounds* are one of the most abundant entities created by nature and they are key components of information space for several technological areas such as drug-discovery (Blundell et al., 2006; Drews, 2000), chemical, agricultural and biofuel research. These *compounds* come as add-ons to macromolecular structures either during their creation or later on at various steps ranging from

DOI: 10.4018/jswis.2010070102

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

sample preparation to crystallization. Building an ontology for *compounds* is a major challenge for developers of structural databases such as the Protein Data Bank (Berman et al., 2000; Berman, Westbrook, Feng, et al., 2000) and the HIV structural database (Prasanna, Vondrasek, Wlodawer, & Bhat, 2005).

Here we present a new rule-based automated technique to develop ontology for *compounds*. We also describe its application to structures chosen from three major databases (1) AIDS structural database[7], (2) Protein Data Bank[8], (3) PubChem[9]. Web tools[10] to query and intersect structural data from these databases using this ontology will also be presented. The terms of the ontology are defined using rules operating on the atomic connectivity of *compounds*. Atomic arrangements are invariants for a given compound and therefore the ontology described here may be considered as "formal" (Gruber, 1993) and thus it is subject to machine reasoning. The terms that form the ontology are of the type Object, Classifier and Value and thus they can be used to generate RDF triples (Lassila & Swick, 1999). An ontology developed by this method for about 3 million compounds obtained from PubChem is available for download[11].

2.1 Background

Structural informatics deals with both large molecules such as proteins and small molecules (*compounds*) such as drugs and co-factors. *Compounds* are usually not peptides and therefore they may not be broken into standard amino-acid like fragments and classified using the rules applicable to protein sequences. Without automated and rule-based methods to classify them, the huge volume and large structural variations in *compounds* pose difficulties for Web tools that try to present and compare *compounds* using their fragments in a predictable and orderly fashion. For instance, during the course of the last twenty years, researchers working on developing AIDS drugs have synthesized thousands of *compounds* that bind to HIV protease (Wlodawer & Erickson, 1993) and these *compounds* share many fragments among them. Identifying these shared fragments (also knows as scaffolds) is a challenge for a Webpage that distributes biological and structural data. Here we describe an automated rule-based method to present and compare *compounds* using their structural ontology built on their fragments.

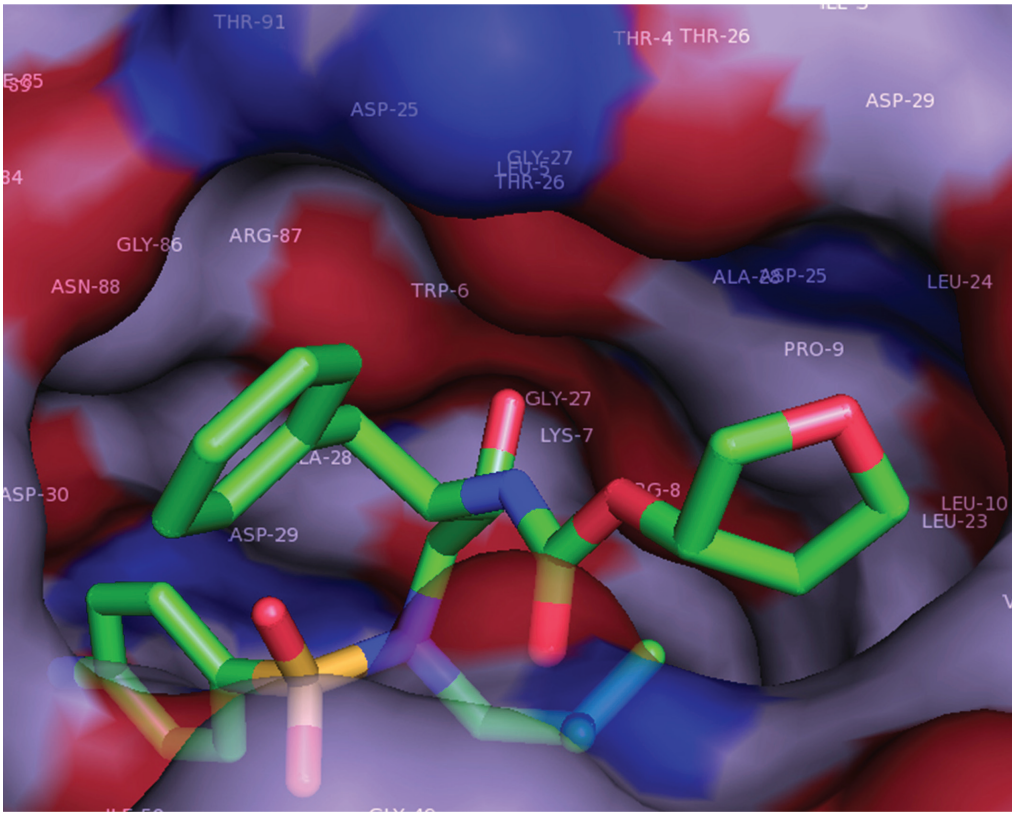
Figure 1 shows the cavity formed by the HIV protease and an AIDS drug, amprenavir, held inside it. This figure is made from the three-dimensional x-ray structure of the protease (PDBID=1YT9) obtained from the Protein Data Bank. The protease cavity is shown as depressions and the grey colored surface shows the vicinity of carbon atoms of a drug molecule in its bound state. The oxygen atom of the drug molecule binds around the red regions of the cavity. Scaffolds of drug molecules that bind to this cavity tend to be structurally very similar since they all have to provide similar interactions to the protein surface around them. The focus of this paper is the automated development of a structural ontology for such drugs using their scaffolds.

2.2 Basic Concepts of the Rules Used to Define Ontological Terms

Compounds that bind to the active site of a particular protein, such as the HIV protease, can be quite different from one another. However, they possess certain common structural components called – scaffolds. Scaffolds are a part of *compounds* (Figure 2) and they are made up of a set of atoms arranged in a certain fashion. These scaffolds bind to specific pockets (HIV protease is considered to have six pockets denoted as P1, P2, P3, P1', P2' and P3') of the active site (Figure 1) of a protein and therefore, in general, one or more scaffolds are conserved among *compounds* that bind to the same protein. The method that we describe here builds an ontology on *compounds* using their scaffolds (Figure 3) and based on programmable rules that are used to define scaffolds (ontological terms) in terms of their atomic bonds. Ontological terms that are commons to multiple *compounds* are used to classify and compare *compounds* and to establish RDF.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 1. Cavity formed by the HIV protease and an AIDS drug, amprenavir, held inside it



2.3 Automated Generation of Scaffolds

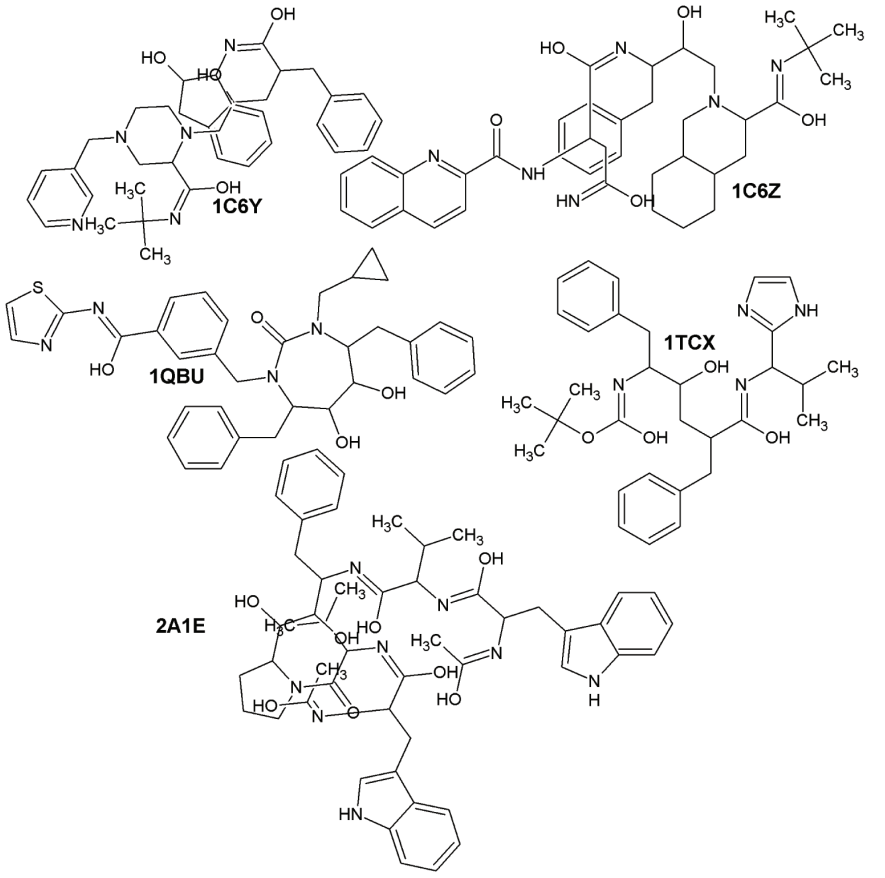
In our method, the scaffolds (Figure 4) are generated by automated procedures working on rules that operate on their atomic connectivity. For instance, a six member ring is coded as a collection of six atoms connected to each other such that starting from any one of these atoms one gets back to the same atom after stepping through six atomic bonds. If any one of those atoms is nitrogen, then it is noted as a nitrogen-containing six member ring. Two rings that share a bond is denoted as a double fused ring. Similar rules are used to define and generate scaffolds of other ring and non-ring scaffolds. These scaffolds are used as elements of RDF triples as described in section 2.4.

2.4 Linking Scaffolds to Form Triples

Triples (Object, Property, Value) of RDF (Table1) are generated by linking scaffolds that have shared structural features. For instance, consider some of the *compounds* (Figure 2) that bind to HIV protease. They have several common structural scaffolds (Figure 3). A six member-ring is one such scaffold. Some of these six member rings have a nitrogen atom in them. Some of the six member rings are fused to another six or five member ring.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 2. Several drug-like chemical compounds that bind to the active site of HIV protease – an AIDS drug target. This paper describes an automated method to build ontology for such compounds by using their scaffolds (Figure 3). PDB Ids are also shown.



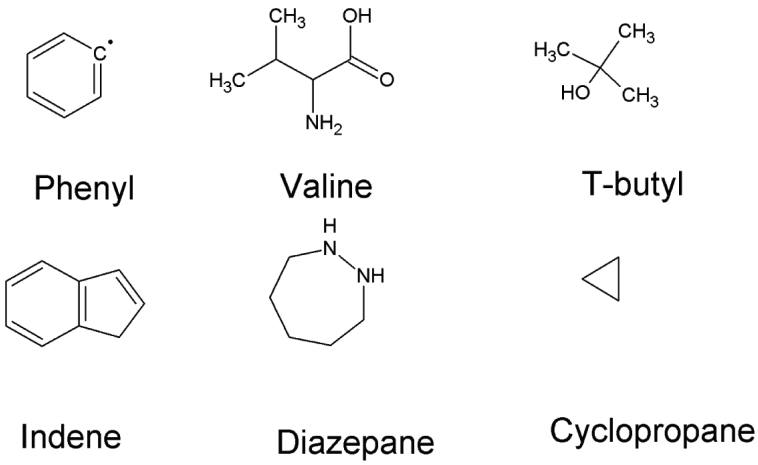
Some *compounds* have a seven member ring. These shared features of scaffolds are used to form their triples as shown in Figure 4. In this table, the middle column lists the Property, and the left and the right columns list Object and Value respectively. IUPAC names of some of these Objects are also shown in Figure 3 and a complete list of IUPAC names can be generated from the InChI described in section 4.1. Though it is plausible to build RDF for scaffolds using IUPAC names, for clarity in recognizing the RDF elements we chose to use images instead.

2.5 Generating RDF Schema

Structural data are too complex to be expressed in a single layer of RDF such as 'Class of' shown in Figure 4. Therefore, a richer multi-layered representation of RDF Schema (Brickley & Guha, 2000) is used. Several layers of RDF are generated using properties called level1, level2,.. leveln. Atoms of an Object in the nth level includes all atoms of its corresponding Object in the $(n-1)^{th}$ level plus the ones in its vicinity. Conceptually, these RDFS gradually

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 3. Some of the scaffolds used to build triples from the compounds such as those shown inFigure 2. Three types of names are used to denote scaffolds; one, short IUPAC names such as Valine, Indene, and the second are InChI (Table 1). InChI are generated by open source software that encodes the information on the atomic connectivity of the atoms of the scaffolds into strings. The third names are unique identifiers used by our database to identify each scaffold and its image. These are about 10 character long strings and they get assigned by an automated step as new scaffolds are discovered.



build the picture of a more complete *compound* starting from certain basic scaffolds such as the rings shown in Figure 4. Examples of the elements of RDFS are given using images in Figure 5 and using the IUPAC names and InChI in Table 1. In Figure 5 the Objects shown in the left column are a subclass of the ones shown in the corresponding right column. The rules used to generate Objects which contain rings are discussed in section 2.3. Atoms that are not part of a ring are considered to form an Object either if they are adjacent to a ring or if they form scaffolds of special interest to drug-design or molecular modeling. Phosphate and t-butyl are two examples of Objects that are not rings. A complete molecule is always considered as an Object of the RDF and it appears in the last (n^{th}) level of RDF. RDF for a subset of the compounds chosen from the PubChem Website is downloadable[12].

2.5.1 Generating OWL from the Triples

Our use of the RDFS did not involve OWL per se. We convert RDFS into data-tree-based ontology and use it in a relational database. However, it is possible to convert the RDFS into OWL and to support SPARQL type queries on them. An example of the OWL for a limited set of structures may be downloaded[13].

2.6 Generation of Ontology

The above described data model with triples provides a source of precisely defined and closely knitted terms that can be used and exchanged across Web applications as ontology. This technique of generating a data model is called Chem-BLAST (Bhat, 2009a, 2009b; Bhat & Barkley, 2007, 2008; Prasanna, Vondrasek, Wlodawer, & Bhat, 2005; Prasanna,

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 4. Examples of triples generated by the method described here. Structural elements though identified by text strings in our database are shown also as images for clarity.

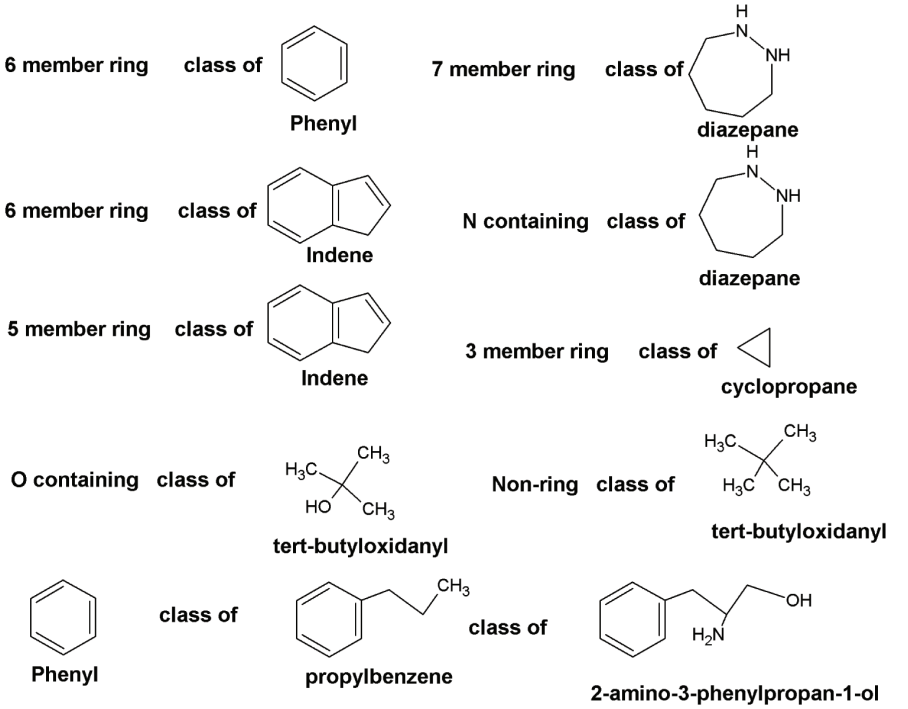
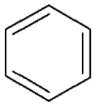
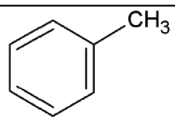
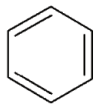
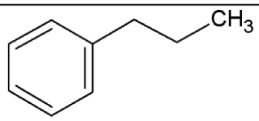
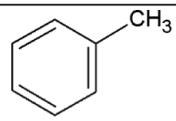
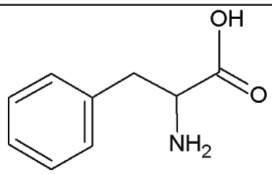
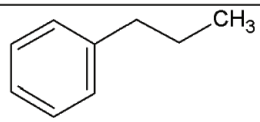


Table 1. Elements of RDFS using text-based names and rule-based names -InChI. InChI is used to assign a unique identifier for each element and to ensure that all identical elements get the same identifier regardless when they are added to the ontology. InChI is also used to re-generate the chemical structure of the elements of RDFS for re-validation purposes.

Phenyl InChI=1/C6H5/c1-2-4-6-5-3-1/h1-5H	level1	Six member ring
Benzyl InChI=1/C7H7/c1-7-5-3-2-4-6-7/h2-6H,1H2	level2	phenyl InChI=1/C6H5/c1-2-4-6-5-3-1/h1-5H
1-phenylethyl InChI=1/C8H9/c1-2-8-6-4-3-5-7-8/h2-7H,1H3	leveln-m	Benzyl InChI=1/C7H7/c1-7-5-3-2-4-6-7/h2-6H,1H2
Phe InChI=1/C9H11NO2/c10-8(9(11)12)6-7-4-2-1-3-5-7/h1-5,8H,6,10H2,(H,11,12)	leveln	1-phenylethyl InChI=1/C8H9/c1-2-8-6-4-3-5-7-8/h2-7H,1H3

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

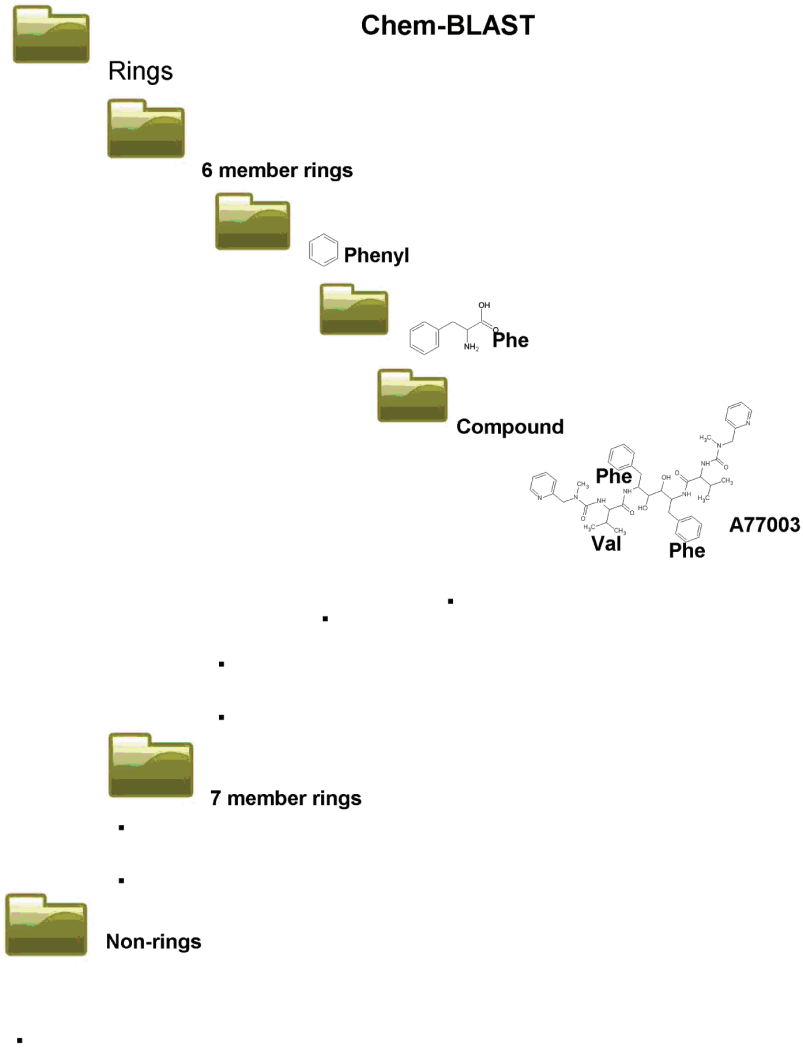
Figure 5. RDFS for a hypothetical compound. The actual implementation of this concept may vary with usage. These elements of RDF all put together form a RDF schema or ontology. In this table elements are shown as images for clarity in specifying as to what they represent.

 benzene	Level1	Six member ring
 methylbenzene	Level2	 benzene
 propylbenzene	Level3	 methylbenzene
 phenylalanine	Level4	 propylbenzene

Vondrasek, Wlodawer, Rodriguez, & Bhat, 2006) – Chemical Block Layered Alignment of Substructure Technique. The code that does this is written in Perl. This code, first extracts all rings from a *compound.* Following that, it examines each ring to look for other rings, if any, around them. Following the identification of all individual and adjacent rings, it extracts non-ring atoms that are adjacent to them, a few atoms at a time. The substructures (rings and non-rings) extracted in these steps are also called scaffolds or fragments. It organizes the scaffolds into groups based on their types such as six or seven member rings. Additionally, it also places scaffolds into an alternative group if any of the atoms they contain, such as Fe, are of ontological interest. Each unique group is considered a 'hub' and a means of association of a hub to another hub through a data model is called 'spoke'. These 'hubs' and 'spokes' collectively form the Objects and Value for building triples and structural ontologies. Keywords such as level1, level2 define the hierarchy (Figure 6) of these Objects. A full *compound* and the type of the ring extracted from it in the first step are the two terminal hubs of the ontology. Additionally, a terminal hub of ontology may be followed or preceded by biological information such as the binding constant of the *compound* or the abstract of the paper that describes the compound.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 6. A schematic view of a structural ontology of a hypothetical compound. The left most layer, called level1, has hubs for the general classification of a scaffold. Next layer, called level2 has hubs for type of the rings it has. The following layer, called level3, has hubs made up of rings and few atoms around them. Actual number of hubs that stack up between the top and bottom layers may vary depending on the size and the chemical structure of a compound.

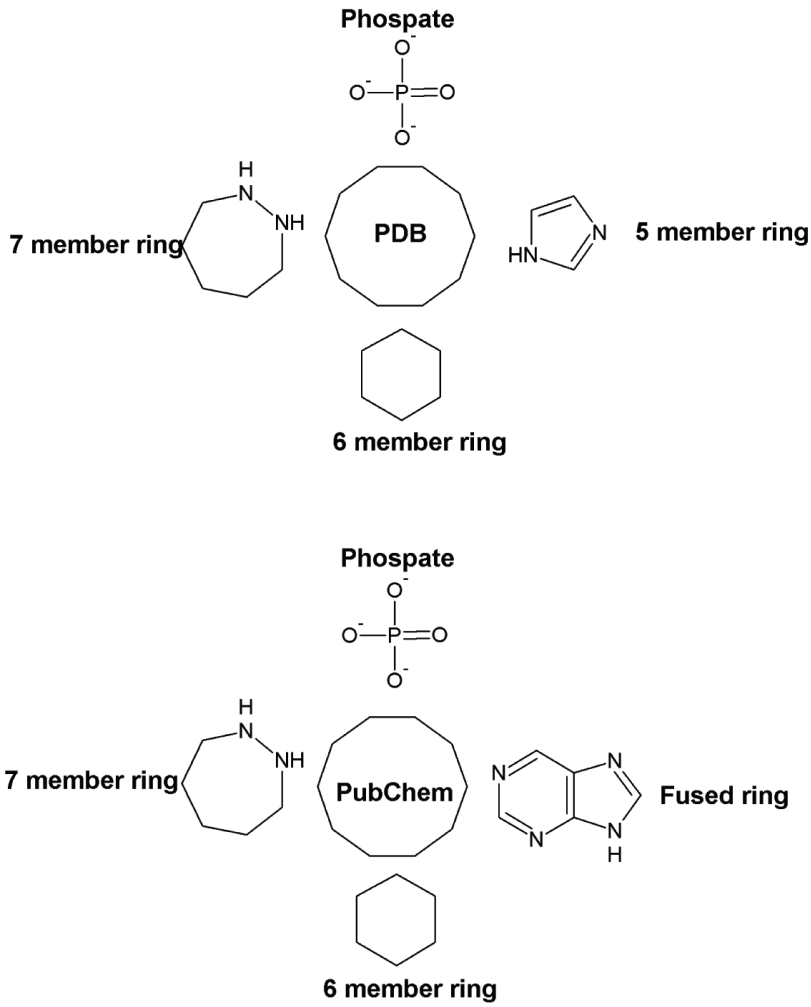


2.6.1 Application of the Method and a Schematic View of the Results

The proposed method was applied to all the *compounds* available in the Protein Data Bank (PDB)[14,15] and also to a subset of the *compounds* in the PubChem Website. A schematic view of the results is shown in Figure 7. This figure shows two layers, an outer layer consisting of Objects e.g., '6 member ring' and an inner layer consisting of Value, e.g., a structure of a benzene ring. Each Object is associated through RDFS to one or more *compounds* in the PDB or in PubChem as the case may be. Figure 7 shows databases as cyclodecane.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 7. A schematic view of some of the results obtained by the method on compounds in the PDB (top) and in a subset of the ones from PubChem (bottom). The method organizes information on each compound into several layers of low to high granularity and connects these layers using RDFS. Elements in these layers and their relationships are used by Chem-BLAST to query and intersect data from these resources.



2.6.2 An Integrated Ontology of All Compounds – Hub-and-Spoke Model of the Elements

An integrated ontology that relates different layers (Figure 7) of information on *compounds* is generated first by collecting distinct InChI of elements of all layers and then by re-naming them so that identical elements have the same name everywhere. The result of this operation is a hub-and-spoke model for the elements of all *compounds* where hubs and spokes are uniquely labeled and classified. In this model similar or identical *compounds* share one or more hubs formed by their scaffolds. Number of hubs shared between two *compounds* depends on the structural similarity between them.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

3. BUILDING A USER-INTERFACE

Although triple store with capabilities to query in SPARQL is a preferred method to query RDF, our interface works differently. Our Web interface deals with RDF elements of structural scaffolds that can be more elegantly presented in a Web as hyperlinked molecular images rather than as concatenated text-strings of their IUPAC names or InChI. Web interface lays out the images in layers of hyperlinked image-tiles for inspection and subsequent selection. Each layer of the image-tile holds images of elements of a particular layer of ontology denoted by the RDF. On selection of a particular image, the Web interface performs a query on the element denoted by that image and displays the result set as new images (Figure 6) for further selection.

The gradual increase in the precision of the image-tiles to define a *compound* in successive layers provides a convenient mechanism to present the images in a predictable order over the Web. The image-tile displayed initially by the Web tool have the scaffolds that point to a collection of both closely and distantly similar *compounds* and the image-tiles displayed subsequently have the scaffolds that point to a collection of *compounds* that are more closely related.

4. ON THE INTEGRATION/ INTERSECTION OF DATABASES USING TRIPLES

We have used *compounds* both from the PDB and PubChem to build a data model using triples and Website to query them. Its schematic view is shown in Figure 8. This Website[16] uses triples to index and integrate/intersect related structures from the PDB and PubChem. Using this Website a user can search similar or identical compounds from the PDB and/or PubChem using the hyperlinks. The hyperlink on a complete *compound* generates list of all entries of that compound in the database.

4.1 Unique IDs for the RDF Elements

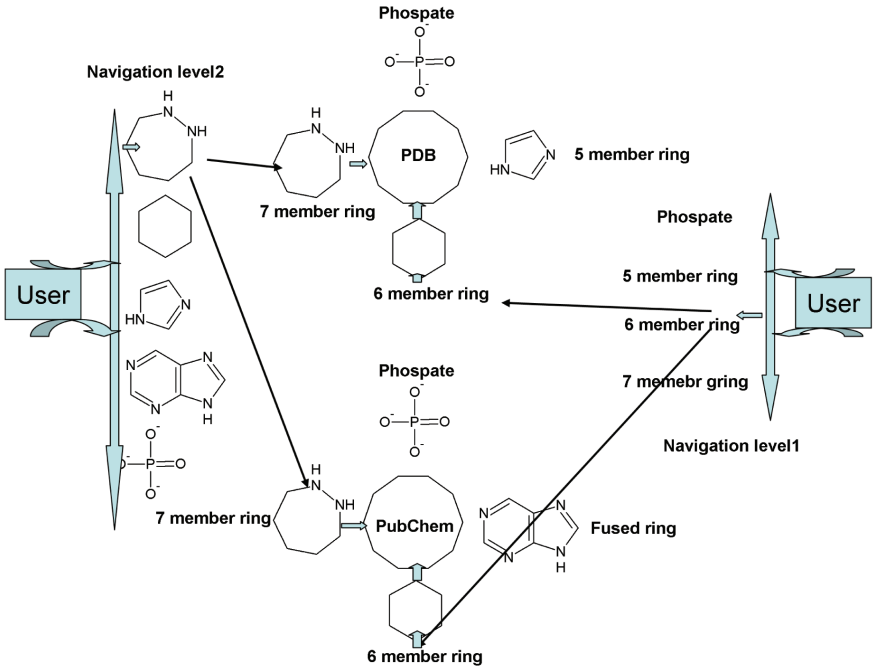
Each element is assigned a unique identifier using the open source software[17]. This identifier is called InChI (International Chemical Identifier). InChI (Murray-Rust, Rzepa, Tyrrell, & Zhang, 2004; Prasanna, Vondrasek, Wlodawer, & Bhat, 2005) is a text string generated from the atomic connectivity of the atoms of the element. The InChI of structurally identical elements will be identical and therefore, all identical elements receive the same identifier regardless of which *compound* they belong to or when they were added to the ontology. InChI encodes the entire atomic connectivity of the element that it identifies, and thus, it may have several hundred characters. Therefore, although InChI provides a convenient method to establish unique identifiers for the elements, it is not suitable as a name of a file that holds the atomic connectivity or image of a scaffold that it identifies. For this reason, a short (usually, less that 10 character long) identifier is also assigned to each element using a look-up table. The current version of InChI is unable to properly handle metal coordination and therefore, a separate identifier may be also needed for accurate representations of *compounds* with metals.

5. RESULTS AND DISCUSSION

Chem-BLAST was initially developed to build ontology for *compounds* of interest to AIDS[18]. As the paper was being written, we extended the work to about 3 million compounds chosen from the PDB and from PubChem. Since the results of the work on the PDB and PubChem have been already discussed in previous sections of this paper, in this section we intend to focus mostly on the *compounds* of interest to AIDS. These *compounds* are traditionally classified on the basis of the mechanism by which they prevent the replication of the AIDS virus. Some of these classes are HIV Protease, HIV Integrate, HIV Reverse Transcriptase, and CCR 5. A few thousand compounds from these

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 8. A schematic view of a user interface to query either the PDB or the PubChem Websites using the RDF elements (shown as images) that are established by the Chem-BLAST. In this interface a user clicks on a molecular image to query for structures either in the PDB or in the PubChem Website. These RDF can be maintained independent of the data resources such the PDB and the PubChem Websites. Therefore, they provide a convenient mechanism to query and intersect federated structural resources without creating a local copy of those resources. This mechanism may help to bring the vision of Semantic Web for structural resources a step closer to reality.



classes were used in our work. Figure 9 shows the Web interface that was developed to query *compounds* belonging to one of these classes. In that figure, the left panel displays information from the top layer of the ontology and this layer holds text-only type information such as amino-acids, and double-fused rings. In the right panel, information on scaffolds from the middle layers is displayed as images.

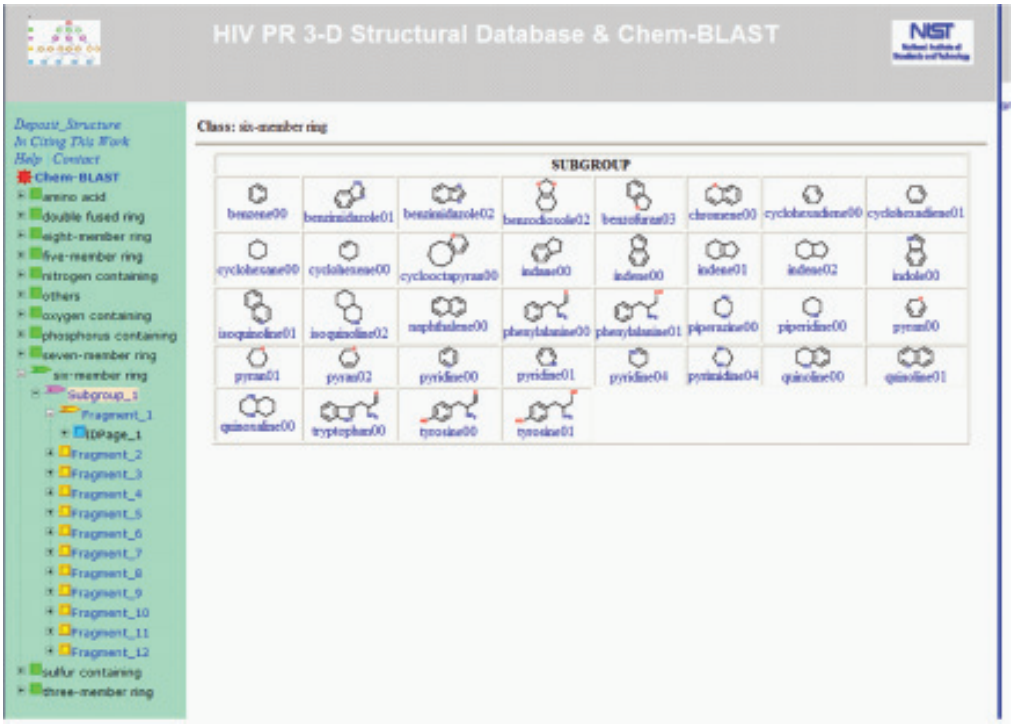
Each data item displayed by the Webpage is a hub for one or more *compounds*. Each hub is hyperlinked to perform query to produce results from its adjacent hubs. The default direction of propagation of a query using a hyperlink is towards its lower level in the ontological tree. Some of the hyperlinks work differently and they propagate query in the reverse (lower to higher) direction and they are shown explicitly using arrows (Figure 10).

Though we store the ontology in a relational database, it can be stored [19] in OWL or RDF using adjacent hubs of the ontological tree as elements and their relationships as predicates.

Chemical *compounds* are an important part of information technology and there have been several major efforts and publications[20,21,22](Batchelor, 2008) on developing their ontology. The primary focus of many

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 9. A part of the Web interface of the AIDS database. It displays the text-based information as folders (left panel) and substructure as image-tiles (right panel). Each Image is hyperlinked to provide query on that scaffold.



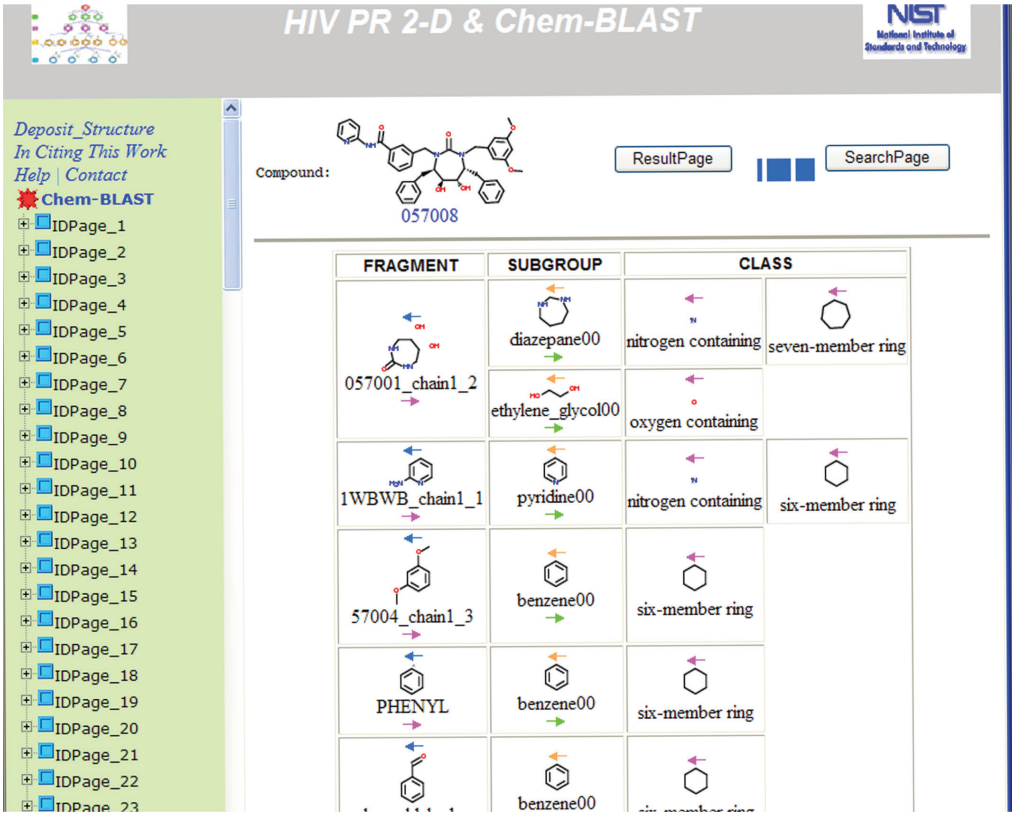
of those efforts has been to develop ontology using a complete *compound*. Though, the work described here has a component of developing ontology for complete *compounds*, its primary focus is to develop ontology for *compounds* using their sub-components (scaffolds) by an automated and rule-based method. The proposed ontology on scaffolds may be integrated with other ontology using either a *compound* or its scaffolds such as the metals and phosphates as the point of intersection between them. A full discussion on published efforts to develop chemical ontology on full *compounds* per se is beyond the scope of this paper.

One of the fundamental properties of a class of *compounds* that bind to a particular target protein, for example the HIV Protease is the reuse of certain scaffolds. This property of a class of *compounds* arises from the fact that all those *compounds* fill the same pocket of the protein (Figure 1). For this reason, we believe that the ontology on scaffolds built by the technique, Chem-BLAST that we describe here has direct applications in chemi-informatics and drug design.

The scaffolds that we used to build the ontology are generated in several levels. Scaffolds at the first level are smaller than the one in its higher level. Scaffolds at the higher levels have atoms from the immediate lower level plus the atoms around them. For instance, (see Figure 5) one of the scaffolds at the first level is a six member ring. The next level of its ontological tree is a six member ring with an additional carbon atom. The ontological tree of scaffolds is presented over a Web as clickable folders (see Figure 9) (Prasanna et al., 2006)[23] . At the first level, they allow a user to select hubs of general descriptions such as six member rings and at the next levels they allow a user to gradually

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 10. A snapshot from the Webpage of HIVSDB that provides hyperlinks to query both in the forward (top to bottom of the ontology tree) and the reverse direction. A user may use a hyperlink in the forward direction to query on spokes to go to its lower layers. Similarly, the reverse arrows provide queries on spokes that connect to its higher layer. The reverse direction hyperlinks result in compounds with lower similarity as it uses a smaller subset of atoms of the structure to query for similar structures. Similarly, the forward direction hyperlinks result in compounds with higher similarity among them.



wean through hubs with more precise definitions of a *compound*.

5.1 Naming of Scaffolds and the Generation of RDF

Atomic bonds between atoms of a scaffold provide a unique mechanism to develop a rule-based method, InChI (McNaught, 2003), for naming them. For instance, the InChI (International Chemical Identifier) for indane molecule (Figure 3) is InChI=1S/C9H10/c1-2-5-9-7-3-6-8(9)4-1/h1-2,4-5H,3,6-7H2. The first part (between the first two "/ ") of an InChI is the chemical formula of the *compound* and the rest is a coded form of its atomic connectivity. It is possible to re-generate the structure of a *compound* (in this case indane) from its InChI using the InChI software[24]. For this reason, InChI provides a novel, rule-based method for naming *compounds* and their scaffolds. However, the InChI does not allow humans to instantly infer the structure of a scaffold. For this reason we developed Chem-BLAST which uses molecular

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

images to scaffolds. Chem-BLAST identifies each scaffold as a file of its image. InChI is not convenient to be used as file names. Therefore, we assign a shorter text-based identifier to identify its molecular image and bond connectivity file. Using this identifier, we generate an RDF (Resource description framework[25]) (Wang, Gorlitsky, & Almeida, 2005) as *compound* – has – *scaffold A; scaffold A – has – scaffold A1; scaffold A – has – scaffold A2* . Elements of an RDF are joined head-to-tail to generate ontology as *compound -> scaffold A -> scaffold A1; compound ->scaffold A ->scaffold A2.*

5.2 Integration of Ontologies

InChI (InChIKey or an alternate shorter identifier described in 4.1) encodes the atomic connectivity of the entire *compound* and thus it may be used to create an invariant URI for a *compound.* However, an integration of information using InChI particularly of a scaffold though can be technically done its ontological relevance to a focus area is less guaranteed. Relevance of a scaffold to a particular focus area such as drug-discovery or patenting depends on the rules used to define it. Moreover, even within a particular focus area for example drug discovery, relevance may be affected depending upon whether it is drug discovery for AIDS or cancer or a target protein. For this reason, a scaffold and its InChI may be treated instead as semi-invariant and used in an ontological URI (OURI) ((Bhat & Barkley, 2007, 2008). OURI may foster 'dialects' with less than perfect overlaps in their meaning in different focus areas and they get their designed meaning only in the context they are designed for.

Compound identifiers used by different databases are often not the same and therefore, the intersection of ontology on *compounds* held by them using their identifiers is a major problem. For instance, the PDB uses what is called a three-letter-code and PubChem uses what is called CID to identify *compounds*. The method proposed here attempts to alleviate this problem by proposing a rule-based system (Bhat, 2009a, 2009b) for generating, and naming elements to build ontology on *compounds* and their scaffolds. Description of a *compound* and scaffolds in the Web is another problem. Chemical structures are too complex to be recognized by their commercial or IUPAC names. For this reason in our Webpage ontological terms, namely scaffolds and *compounds*, are stored and displayed as images. However, from their identifiers descriptive IUPAC names could have been generated for ontological terms and a descriptive RDF could have been also built (Figure 3). Such an RDF would have been of the type *Indane – has – Six member ring* and *Indane – has – four member ring. Compound 006688 – has – Indane.* Additional examples of this type of RDF have been posted[26] in a downloadable OWL document[27]. In our application, we generate the ontology as triples and store them in a relational database. However, these triples could have been stored in a triple store and queried using SPARQL as well. After we started to write this paper, we extended our work to few million compounds chosen from the PubChem Webpage. A part of its result is downloadable[28]. This result is also available to the community as a standard reference data distributed by NIST.

ACKNOWLEDGMENTS

Work on HIVSDB was done in part in collaboration with Anh-Dao Nguyen of NIST, G. Noble and L. Cooney of Cygnus Corporation, M. Nasr of NIAID, A. Wlodawer of NCI and K. Das and E. Arnold of University of Rutgers.

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.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., & Miller, W. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. doi:10.1093/nar/25.17.3389

Batchelor, C. (2008). Formal ontology in information systems. In *Frontiers in Artificial Intelligence and Applications* (Vol. 183, pp. 195-207).

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., & Weissig, H. (2000). The PDB and the Challenge of Structural Genomics. *Nature Structural Biology*, *11*, 957–959. doi:10.1038/80734

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., & Weissig, H. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*, 235–242. doi:10.1093/nar/28.1.235

Bhat, T. N. (2009a). Chemical Taxonomies and Ontologies for Semantic Web. *Semantic Universe.* Retrieved from http://www.semanticuniverse.com/articles-chemical-taxonomies-and-ontologies-semantic-web.html

Bhat, T. N. (2009b). *On the creation of structural FaceBook using rule-based methods to build and exchange ontology for drug design.* Retrieved from http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-549

Bhat, T. N., & Barkley, J. (2007). Semantic Web for the Life Sciences - Hype, Why, How and Use Case for AIDS Inhibitors. In *Proceedings of the IEEE Congress on Services* (pp. 87-91).

Bhat, T. N., & Barkley, J. (2008). Development of a Use case for Chemical Resource Description Framework for Acquired Immune Deficiency Syndrome Drug Discovery. *The Open Bioinformatics Journal*, *2*, 20–27. doi:10.2174/1875036200802010020

Blundell, T. L., Sibanda, B. L., Montalvao, R. W., Brewerton, S., Chelliah, V., & Worth, C. L. (2006). Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philosophical Transactions of the Royal Society B . Biological Sciences*, *361*(1467), 413–423. doi:10.1098/rstb.2005.1800

Brickley, D., & Guha, R. V. (2000). *Resource Description Frame Work (RDF) Schema Specification 1.0, W3C Candidate Recommendation.* Retrieved from http://www.w3.org/TR/rdf-schema

Drews, J. (2000). Drug discovery: a historical perspective. *Science*, *287*, 1960–1964. doi:10.1126/science.287.5460.1960

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition*, *5*, 199–220. doi:10.1006/knac.1993.1008

Lassila, O., & Swick, R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation*. Retrieved from http://www.w3.org/TR/REC-rdf-syntax

McNaught, A. (2003). *What's in a Name*. The Alchemist.

Murray-Rust, P., Rzepa, H. S., Tyrrell, S. M., & Zhang, Y. (2004). Representation and use of chemistry in the global electronic age. *Organic & Biomolecular Chemistry*, *2*(22), 3192–3203. doi:10.1039/b410732b

Prasanna, M., Vondrasek, J., Wlodawer, A., & Bhat, T. N. (2005). Application of InChI to curate, index and query 3-D structures. *Proteins, Structure, Function, and Bioinformatics*, *60*, 1–4. doi:10.1002/prot.20469

Prasanna, M. D., Vondrasek, J., Wlodawer, A., Rodriguez, H., & Bhat, T. N. (2006). Chemical compound navigator: a web-based chem-BLAST, chemical taxonomy-based search engine for browsing compounds. *Proteins*, *63*(4), 907–917. doi:10.1002/prot.20914

Wang, X., Gorlitsky, R., & Almeida, J. S. (2005). From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nature Biotechnology*, *23*(9), 1099–1103. doi:10.1038/nbt1139

Wlodawer, A., & Erickson, J. W. (1993). Structure-based inhibitors of HIV-1 protease. *Annual Review of Biochemistry*, *62*, 543–585. doi:10.1146/annurev.bi.62.070193.002551

ENDNOTES

1. http://bioinfo.nist.gov/SemanticWeb_pr2d/chemblast.do
2. http://www.rcsb.org/pdb/explore/externalReferences.do?structureId=3GGT
3. http://xpdb.nist.gov/chemblast/pdb.html
4. http://xpdb.nist.gov/chemblast/chemlevel1.pl?T1=IMD
5. http://xpdb.nist.gov/pubchem_chem_blast/rdf/pubchem.rdf
6. http://xpdb.nist.gov/hiv2_d/download.html

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

7 http://bioinfo.nist.gov/SemanticWeb_pr2d/chemblast.do
8 http://www.pdb.org/pdb/home/home.do
9 http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=00001199
10 http://xpdb.nist.gov/chemblast/pdb.pl?pdbid=3GGT
11 http://xpdb.nist.gov/pubchem_chem_blast/rdf/pubchem.rdf
12 http://xpdb.nist.gov/pubchem_chem_blast/rdf/pubchem.rdf
13 http://xpdb.nist.gov/hiv2_d/download.html (The Properties used here are subject to change).
14 http://www.pdb.org/pdb/home/home.do
15 http://pubchem.ncbi.nlm.nih.gov/
16 http://xpdb.nist.gov/chemblast/pdb.pl?pdbid=3GGT
17 http://www.iupac.org/inchi/
18 http://bioinfo.nist.gov/SemanticWeb_pr2d/chemblast.do
19 http://esw.w3.org/topic/HCLS/ChemicalTaxonomiesUseCase
20 http://www.semanticuniverse.com/
21 http://pubs.acs.org/doi/abs/10.1021/ci060139e
22 http://www.ebi.ac.uk/chebi/
23 http://bioinfo.nist.gov/SemanticWeb_pr2d/chemblast.do
24 http://www.iupac.org/inchi/
25 http://www.w3.org/RDF/
26 http://esw.w3.org/topic/HCLS/ChemicalTaxonomiesUseCase
27 http://xpdb.nist.gov/hiv2_d/download.ht
28 http://xpdb.nist.gov/chemblast/chemlevel1.pl?T1=1T3R_l1_2

Talapady N. Bhat is a Project leader in the Biochemical Science Division of the Material Measurement Laboratory of the National Institute of Standards and Technology. He is one of the authors of over 80 refereed scientific articles that have so far received over 13,000 scientific citations (source ISI Web of Knowledge). Dr. Bhat has won several scientific awards including: two Best Publication Awards from the Science Applications International Corporation; the Judson C. French award of the National Institute of Standards and Technology in 2005; the Science Spectrum Traiblazer Award and the Emerald Honors from Science Spectrum Magazine in 2006 and 2007 respectively.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.