Semantic Web for the Life Sciences - Hype, Why, How and Use Case for AIDS Inhibitors

Bhat, T. Narayana
Biochemical Science Division, NIST, Gaithersburg, MD 20899 USA
bhat@nist.gov Ph: 301 975 5448

Barkley, John
Software Diagnostics and Conformation Testing Division, NIST, MD 20899 USA
Jbarkley@nist.gov

Abstract

One of the difficulties in applying the concepts of Semantic Web technology to a drug discovery is the creation of a software generated URI suitable both for ontology independent and ontology dependent properties of drugs. We propose the use of two types of URI: structure invariants (URI) to identify the ontology independent features and structure semi-invariants (OURI) to identify ontology dependent features. URI are defined by the most fundamental properties (chemical connectivity, bond type, etc.). OURI are defined from context dependent (semantic) properties (binding mode, chemical core)

1. Introduction

There has been considerable effort to establish (e.g. YeastHub[1] BioPax - http://www.biopax.org/, ChEBI - http://www.ebi.ac.uk/chebi/, SMID http://s Wikimedia.org/snid_about.php ) the Semantic Web as a vision for the future of Web technology for biological sciences. As a part of the efforts by the Health Care and Life Sciences (HCLS) Special Interest Group (SIG http://www.w3.org/2001/sw/hcls/), we have been developing a use case for chemical structures for drug-design purposes for AIDS with the data obtained in four major independently maintained Web browsers (the Protein Databank - PDB[2] , HIV Structural Database - HIVSDB[3, 4], AIDSD (http://chemdb2.niaid.nih.gov/struct_search/defa ult.asp), and PubChem[5]).

Biological applications and the drug discovery processes is complex. Scientists build precise model using aggregates of subset of a structure. Scientists would like to use such subset of structures to exchange information over the Web, for instance, to obtain structurally similar (similarity defined in ontology) drugs from a collection of drugs. The need for the exchange of information over the Web for a subset of a drug using an URI creates practical difficulties – how to specify the structural information about a subset of a drug using an identifier without compromising its context with respect to other subsets of the same drug that collectively create the model of interest to a use case? In other words, the question is - how to assign a URI to a subset of a drug such that such URI is also unambiguously associated with URI of another subset of the same drug that in combination represents the view of the use case? Another question is how to define an ontologically defined subset using a non-ontologically defined identifier?

We propose the use of two types of URIs, a context invariant URI that may be used to exchange information between resources devoid of information on a particular use case; the other, a context dependent URI (semi-invariant URI, ontologically defined URI (OURI)) that may be used to exchange information only in a given use case defined by a taxonomy/ontology. How an OURI is created is dependent on the commonly
asked questions by users of a particular use case. In this paper, we show how these OURIs may be defined and used in a Web for AIDS inhibitors.

2. Drug design cycle

For the sake of completeness, we would like to present an overview of a drug design cycle and to clarify some of the concepts we used to develop a use cases for the Semantic Web. A drug design cycle has two main components, a) in silico experiments; b) in lab synthesis and testing. In silico experiments are done by structure biologists and modelers and these experiments propose new drug candidates by comparing the existing drugs in view of some hypothesis. The lab synthesis step carries out the chemical reaction to make the chemical proposed by a modeling step.

3. Definition of an invariant URI

Numerous techniques have been proposed and used to name chemical structures. Some of these (CAS numbers, PDB 3-letter codes, commercial names) provide a unique identifier; but unfortunately they are vocabulary-based, and thus they are not amenable to machine reasoning. IUPAC names are generated automatically using certain rules. However, the rules used in their assignments are not robust enough to produce an unique identifier[3] for a compound. For this reason, we propose to use the recently announced IUPAC International Chemical Identifier (InChI[3, 6]) for generating the invariant URI for a drug. The InChI has several embedded special characters, such as, “/”, space,” (“,“),and “’”。 In order to make InChI names compatible with Semantic Web tools, InChI names are transformed first by URI encoding, and then, each “%” resulting from the encoding is replaced by “_”.

4. Definition of semi-invariant URI from ontology

A semi-invariant URI (OURI) of a compound defines one of the many ontologically defined subsets of the compound identified by a particular invariant URI. The rules of decomposing a compound into smaller subsets for creating an OURI may vary depending on the use case. A molecular modeler may view a compound to be made up of structural subsets that bind to specific pockets of a target enzyme[3, 4, 7]. These pockets of the AIDS protease (fig 1) are defined as S1, S2, S3, S1’, S2’ and S3’. The use case that addresses the needs of a molecular modeler, therefore, may consider AIDS drug to be made up of subsets that are specific to each of these sub-sites of the enzyme. A molecular modeler would be interested in comparing (queries using concepts - union, intersection and complement) drugs on the subsets that bind to these sub-sites. Therefore, we propose the use of OURI to define these ontologically defined subsets. We use InChI and/or commonly used vocabulary (figure 1) to represent the OURI.

Medicinal chemists may have a view different from that of modelers. Medicinal chemists work on the synthesis of a drug using the molecular model proposed by a modeler. They may view the drug as made up of subsets that combine during a chemical reaction to form the drug. A use case that focuses on the needs of a medicinal chemist, therefore, may define a OURI in terms of the chemical fragments from which the drug has been synthesized.

Biologists who do the vitro testing of the drug may have an entirely different view of the drug. They may focus on specific molecular groups, such as polar groups, that affect bio-availability. For this reason, a use case based on chemical synthesis or molecular modeling may not be suitable for a biologist. Similarly, Patent lawyers may have a view (usually based on the chemical core of the drug) that is entirely different from those of others.
Figure 1. Shows the active site of HIV protease - an enzyme that is critical for the replication of AIDS virus. Also shown is a drug bound to its active site. A modeler’s view is that the active site is made up of sub-sites S1, S2, S3, S1’, S2’ and S3’ and he compares drugs by performing intersection and union on the portions of the drug that bind to these sub-sites. To support Semantic Web with these capabilities we propose to define OURI for the part of a drug that bind to these sub-sites.

In spite of all these different requirements the fact is that all the experts involved in a drug development need to exchange information on the same drug. To facilitate this, we propose the use of URI (common to all use cases) in conjunction with the appropriate OURI (specific to a use case).

Figure 2. Shows a chemical taxonomy that organizes the subsets of a drug into a data tree. Elements of the data tree are treated as a semi-invariant URI and the complete drug denoted by the data tree is identified by a URI. A Web browser interprets the semi-invariant URI within the context of the other elements of a taxonomy. Though a semi-invariant URI has only a subset of the complete information on a drug transmitted over the Web, the Web browser resolves it by interpreting its semantics in the context of other elements the taxonomy tree. In this figure (taken from http://xpdb.nist.gov/hiv2_d/advanced_query_files/slide0002.htm) ‘Start here’ marks the hypelinks wherein a user may chose to initiate queries on semi-invariants.

5. Using OURIs

We chose to illustrate the use of developing Semantic Web use case (http://esw.w3.org/topic/HCLS/ChemicalTaxonomiesUseCase) for about 2000 structures obtained from four major Web pages (PDB, HIVSDB, PubChem and AIDSDB) that distribute data on AIDS drugs. Subsets of these structures were assigned OURI from a) structural biology and modeling view, and b) the medicinal chemist’s view that we discussed earlier. (We do not claim to have included all the necessary views in these topics). Synonyms often add ontological values for humans and thus, popular vocabulary based synonyms were used when they add values for the concepts. These OURI may be used to compare[8, 9] drugs using the URL given in fig 2. This comparison may be then used to propose new drug candidates that may have superior qualities compared to others used in the comparison. Drug-resistance is a very common problem for AIDS drugs[8] and this issue has been the primary focus of drug design efforts in recent years. Medicinal chemists may use a different view than modelers. Medicinal chemists use the information generated by structural biology and the modeling view to develop chemically synthesizable and biologically viable drugs.
6. Web Development

The URIs and OURIs are then organized into a chemical taxonomy[4] (fig 2) to reflect, if any, relationship (intersection, union) among them. The information is then expressed in OWL. From the atomic connectivity of structures denoted by each URI/OURI, molecular images were created for the elements of the taxonomy. Web tools were developed to present the elements of the chemical taxonomy using these molecular images. These images were hyperlinked (http://xpdb.nist.gov/hiv2_d/advanced_query_files/slide0002.htm) to facilitate queries on the use case views. In our implementation, we use SQL running on relational databases managed by MySQL/Oracle for storage and retrieval of the data. Additional details on the use of the Semantic concepts to get structural and biological data from the above listed four Web pages may be found at the W3C Web site given above. The OWL for this use case may be downloaded from the W3C Web site. During my talk details of the implementation of the concepts using relational databases such as Oracle 9i or MySQL will be presented.

7. Discussion

Use of the uniform ontological standards is easily said than done. Different use case requirements, practicality, different software requirements, and different annotation techniques are just a few of the factors that work against the use of uniform ontological standards. Here we illustrate how some of the concepts of Semantic Web can be implemented for chemical structures using an evolutionary approach involving a set of global and local identifiers. We sub-divide the URI into globally accepted URI and locally implemented ontological OURI. We illustrate the proof of this concept using data from four independently maintained databases (PDB, HIVSDB, AIDSDB and PubChem) that have complementary information on AIDS drugs. OURI are handled only by local search engines. Once a query on OURI hits a URI that is accepted by other participating Web resources, local search engines pass on control (http://xpdb.nist.gov/hiv2_d/hivsdb_fragment_search.pl ) to relevant external Web resource. The Web resource that receives control (AIDSDB) then deploys its own ontology and continues with new options for users and pass control back to the parent Web page (HIVSDB) as illustrated in (http://chemdb2.niaid.nih.gov/struct_search/misc/url_search.asp?aids_no=000709)

For decades, researchers have been developing controlled vocabulary for naming substances. These vocabularies have had a major impact in the way researchers communicate information. Indeed, they are the best known method to exchange information using a centralized repertoire of names that may be meaningful for humans. However, this method is ill suited for assigning URI in a modern Web based global information exchange systems. The Web has created a need for a system that can be used independently and simultaneously by machines to uniquely name substances (or subsets of a substance) that may or may not exist in the existing vocabulary. A rule-based URI combined with vocabulary and/or the ontology based OURI proposed here presents an alternative approach that facilitates the co-existence of both global and local standards during Semantic Web developments.

8. 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.

9. References


