Chapter 1
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; (b) ~60,000 structures from the PDB. Recently, this technique has been applied to approximately 3,000,000 compounds from PubChem. Plausible ways to use this data model for the Semantic Web are also discussed.

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

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 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, (2) Protein Data Bank, (3) PubChem. Web tools 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.

2. 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 known 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