

FGO: A novel ontology for identification of ligand functional group

Prithish Kumar Varadwaj¹ and Tapobrata Lahiri^{1*}

¹Indian Institute of Information Technology, Allahabad, India - 211012; Tapobrata Lahiri * - E-mail: tlahiri@iiita.ac.in;

* Corresponding author

received July 17, 2007; revised November 03, 2007; accepted November 06, 2007; published online December 05, 2007

Abstract:

Small molecules play crucial role in the modulation of biological functions by interacting with specific macromolecules. Hence small molecule interactions are captured by a variety of experimental methods to estimate and propose correlations between molecular structures to their biological activities. The tremendous expanse in publicly available small molecules is also driving new efforts to better understand interactions involving small molecules particularly in area of drug docking and pharmacogenomics. We have studied and designed a functional group identification system with the associated ontology for it. The functional group identification system can detect the functional group components from given ligand structure with specific coordinate information. Functional group ontology (FGO) proposed by us is a structured classification of chemical functional group which acts as an important source of prior knowledge that may be automatically integrated to support identification, categorization and predictive data analysis tasks. We have used a new annotation method which can be used to construct the original structure from given ontological expression using exact coordinate information. Here, we also discuss about ontology-driven similarity measure of functional groups and uses of such novel ontology for pharmacophore searching and de-novo ligand designing.

Keywords: functional group; ontology; knowledgebase; semantic similarity; data mining; database; pharmacophore

Background:

The concept of ontology as an abstraction from the actual data entities and instances is not a new concept. Philosophers have been studying the theory of objects and their ties for centuries. However, ontology, as we know them today has become more formalized and conceptual in computer science, database integration, Bioinformatics and artificial intelligence. [1] Ontology of biological terminology known as Bio-ontology, such as the Gene Ontology or Chemical Ontology provides a model of biological concepts that can be used to form a semantic framework for many data storage, retrieval and analysis tasks. [2] Such a semantic framework could be used to underpin a range of important bioinformatics tasks, such as querying of heterogeneous bioinformatics sources or the systematic annotation of experimental result. [3] Ontology is designed to define concepts that are necessary to describe the domain but not individual instance in that domain the controlled vocabularies, which have been proved to be extremely useful to researchers, in order to aid in the classification and organization of information. [4] In arena of biological and chemical ontology we have few existing systems, each with its advantages and limitations. For example, gene ontology (GO) [5] is used to describe and classify information regarding protein function at the same time as it also help in grouping together similar objects, or find things with similar properties or behaviors. Hence this ontology is highly used to facilitate data exchange, ISSN 0973-2063

analysis, and searching. Similarly biochemical ontology, BAO has been proposed using Powerloom, which is claimed to be more expressive, flexible and extendible than the traditional structures used before. [4] But very less attempt has been made for a suitable small molecule ontology to classify, represent and search available small molecule and its interactions. Though small molecule and its interaction are highly essential in field of drug designing and pharmacogenomics and due to the constant increments in availability of such data a formal chemical ontology for describing small molecules is of strong need. With exception of Chemical and Biological Interaction (ChEBI) [6], and Chemical Ontology (CO) [7] no other options are available till date. ChEBI hosted at the European Bioinformatics Institute (EBI), developed ontology to help classify small molecules in their database. Each entry in the database is manually added as a leaf along one or more branches in the ontological tree. ChEBI classify with many levels of specificity within the ontological tree, and it does so with both by chemistry and by molecular function. But major drawback is every time a new compound is added to the ChEBI database, it must be assigned to the ontological tree by hand, which is obviously very labor intensive and cumbersome. It is also somewhat subjective because of many such new terms themselves are somewhat vague and open to interpretation, since they are not defined strictly. ChEBI terms may have multiple parents hence it is difficult

to establish and maintain relationships in a growing ontology. However, they can be better handled by formal expression and the reasoning capabilities of underlying description logic such as DAML + OIL [5], CO [7] is another available solution where the small molecules were defined by the constituent functional groups. It's a well established fact that each ligand can be considered as a graph comprising of largely independent sub graphs i.e. chemical functional groups. So a classification for ligand could also be obtained by identifying and arranging such cliques. As well a hierarchical approach of such functional groups classification is desirable in maintaining simplicity, allowing simple searching and ontological assignment while providing a set of rich descriptive terms that can be used for semantic comparison. CO addresses the automatic detection of functional group consistently and objectively assigned by a FORTAN computer program Checkmol. [8] It has also proposed a semantic similarity metric similar to the Tanimoto score [9]. But the major drawback with CO is the unavailability of geometric orientation information of ligand atoms. A minor change in ligand conformation can change the overall binding affinity with receptor molecule; hence an effective way of representing coordinate information should be included in chemical ontology. The portion of a small molecule responsible for molecular recognition or binding in a particular active site is referred as a pharmacophore. This could be further defined as the geometrical arrangement of specific chemical moieties of ligand molecule to bind and produce pharmacological actions on receptor molecule. These can be numbers of hydrogen bond donors/acceptors in the binding site, charges if any, aromatic ring centers, and so forth. So ligand molecules should be compared based on the presence of potential pharmacophore and the similarity between such pharmacophore. Recently, there has been an increased effort in developing methodologies for the design and evaluation of ontologies that can address the above challenges. The Functional Group Ontology (FGO) proposed by us tries to incorporate the flexibility of above described Ontologies and on other hand has uniquely overcome with the limitations discussed with previous adopted methodology. We have included the geometric orientation profile information of ligand atoms and also given a simple annotation key for deriving semantic similarity. Additionally, it can be used as a powerful search interface, allowing the pharmacophore searching with any desired combination of functional groups. It has also been integrated with the in-house ligand repository of 11800 ligand structure and also been used in representing the ligand atom in protein-ligand complex PDB file obtained from. [10] Basic design criteria for a knowledge representation system include integrity, consistency, and clarity. [11, 12] In addition, while examining the biological and chemical domain following four new design criteria were identified granularity, abstraction, independence, and isolation. FGO tries to include these four criteria to make it robust and reusable. The functional group ontology described in this paper use hierarchical structures for

objects and properties as well as a network of relations hence results in a finer granularity, and enhances the isolation of the concepts.

Methodology:

We have designed a tool which is able to automatically detect and assign functional group (FG) information of any given small molecule. Practically it is able to identify any functional group though only 148 functional groups have been annotated in our study. Assignment of FG is made strictly based on computational detection of specific arrangement of atoms and bond with in the input molecular structure. A given ligand structure may have any number of FGs assigned to it and the detection of these FG were carried out with specific coordinate information. Functional group detection program has been carried out by MATLAB tool. After identifying functional groups for given small molecule we have designed and represented the constituent functional groups by our novel ontology (FGO). We have used OBO-Edit tool for the creation of Functional Group Ontology.

Ligand three dimensional structures (MOL file) were taken as inputs. [13] To extract the functional group from protein-ligand complex, we have taken PDB file as input. Since PDB file do not have connection information we have initially converted PDB files in to MOL file and heteroatom coordinate information of PDB file and its connection information from MOL file were specifically extracted. Coordinate information of water molecules are explicitly removed and only small molecule coordinates information were taken.

Such MOL file connection block information was parsed and adjacency matrix was created for further processing. Each row and column information of above matrix was taken in to consideration to check the highest bond order priority based link matrix data. Corresponding atom information with each bond order was extracted with respective coordinate information, to categorize and identify each probable functional group in given ligand (figure 1). Information was mapped to chemical tree for functional group identification and annotation in descending bond order priority. Annotation key were used in FGO to identify and represent the functional group structural information. Apart from coordinate information we have also considered and encapsulated other relevant structural parameters like, presence of ring, aromaticity, charge and stereochemistry. Above information were represented by FGO, which was generated using OBO-Edit. OBO-Edit is developed by John Day-Richter of the Berkeley Biomedical Ontologies Project. It is an open source ontology editor written in Java and with an easy-to-use editing interface. We have designed and implemented a chemical tree originated by a Carbon atom. Our developed chemical tree is having seven branches at each hierarchical level namely carbon, oxygen, nitrogen, sulphur,

phosphorus, halides and misc abbreviated for other atom types.

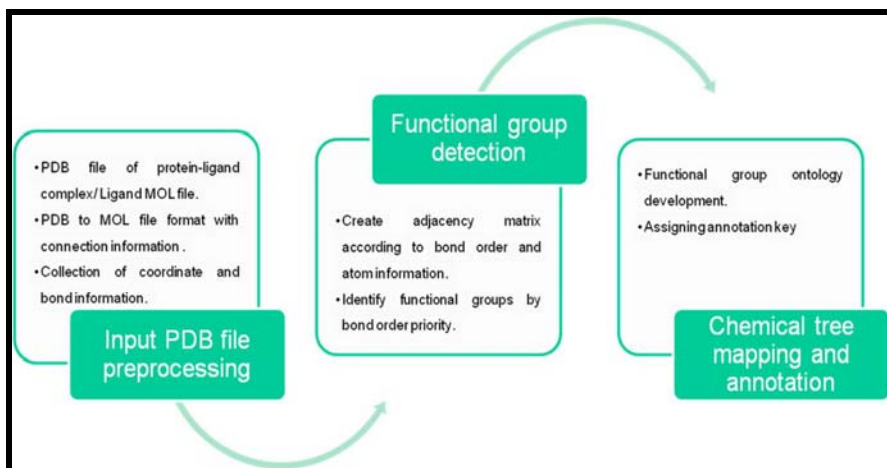


Figure 1: Flow chart of functional group ontology generation and annotation key assignment

The atoms other than the list and all metal atoms were categorized in misc. branch. We have four connective instances for each level of branching namely single, double, triple and special for bond type consideration (figure 2). Partial bond, resonance, distributed electronic clouds and other than covalent bonds are treated as special case. Each branch is having a fixed class (figure 3) and also gives rise to another set of seven branches with exception of leaf

nodes, which has no branches. Each class is having fixed architecture carrying all information about the atom in consideration and cumulatively each ligand functional group is a specific topology of branches with set of connected nodes. We have generated IDs by creating new child and 20 level of such branching has been designed to specially accommodate bulkier steroidal parent chain.

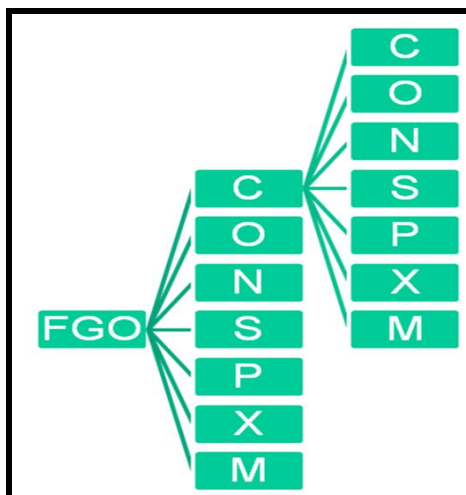


Figure 2: Chemical tree with single bond order instance and seven branches

The ligand information were retrieved and parsed through the chemical tree with specific path and topology to represent them uniquely with specific annotation for each step. Each new child node has been annotated by atom type and bond order information. The hierarchical chemical tree generated has been designed to carry coordinate

information of each node connected to it, which takes the atomic orientation profile into account. The FGO has been designed to treat each functional group as an abstract super class with each atom of it as an object. The detailed information of each atom and its connection information with parent atoms were stored in a single class.

```

atom_type
{
  connection_type
  atom_type
  {
    Connection_type
    atom_type
    coordinate_information
    ring_information
    extra_information[dynamic_list]
    annotation_key}
  coordinate_information
  ring_information
  extra_information[dynamic_list]
  annotation_key}

```

Figure 3: Class architecture of a single node

Such classes were hierarchically arranged with specific annotation to denote a specific set of instances, hence a functional group.

Annotation keys

For annotating a functional group each of the constituent node instances were considered with the atomic and connectivity information. We have used an annotation rule similar to single line entry used to denote molecular structure. Atom information of all seven types were denoted as C, N, O, S, P, X, M for carbon, nitrogen, oxygen, sulphur, phosphorus, halide and others respectively. Similarly, four connectivity instances were taken as 1, 2, 3, 4 for single, double, triple and others, respectively. Chiral and aromatic properties were taken care by similar treatment like SMILES. [14] We have used @, / and above set of lower case alphabets for representing chirality, E and Z type isomerism and aromaticity of specific atom and functional group respectively. Branching in structure was represented by starting and closing round brackets, with concatenation of two specific single chain annotations (Table 1 in supplementary material). The ring structures were annotated by curly brackets. Though hydrogen information is missing in MOL file we have not considered hydrogen information and its role has been assumed to satisfy the valance of other atoms.

Semantic similarity

Let S be the set of terms used in the FGO (functional group ontology). Information-theoretic approaches to measuring similarity between terms, $s \in S$, may be based on the amount of information associated with them or shared by them in common. Given a pair of ligand, L_i and L_j , which are annotated by a set of terms A_i and A_j respectively, where A_i and A_j comprise m and n terms respectively, the semantic similarity, $SIM(L_i, L_j)$, may be defined as the average inter-set similarity between terms from A_i and A_j ; and $sim(s_i, s_j)$ represent the similarity between terms as given in equation

ISSN 0973-2063

116

Bioinformatics 2(3): 113-118 (2007)

1 (see supplementary material). The semantic score can be between 0 to 1, showing no similarity and perfect similarity respectively.

Result and discussion:

The expressive representation of ontology is to extract information and hence decision making from set of abstract data. The description logics used in ontology are knowledge representation languages tailored for expressing knowledge about concepts and concept hierarchies. We are assessing and detecting functional group of any ligand on the basis of their atom types, bond information and other structural properties of ligand molecule. The ligand functional group detection program has detected 148 different functional groups, which are useful for representing any specific pharmacophore. It has been applied to in-house data base of 11800 ligand structures to demonstrate its utility as a basic pharmacophore search system, ligand classifier and similarity based clustering measure. Molecules have been compared using a semantic similarity score based on functional group which succeeds in identifying small molecules known to bind a common binding site (table 1 in supplementary material).

This ontology will serve as a powerful tool for searching chemical databases and identifying key functional groups responsible for biological activities. The purpose in validating the ontology is to ensure the consistency and reliability of the ontology. This includes the ability to reliably retrieve query paths, valid relationships, and also to conduct complex queries. Analysis and testing were conducted through a number of test queries in the above database and was shown to be faster than graph based sub-structure searching algorithm. Classifying concepts involves the classification of the hierarchical tree relationships between classes. In this case, the inference system has classified the ontology according to its hierarchical structure, i.e. traverse and display super and

sub-classes. Queries were performed upon the ontology in order to test whether the ontology would perform as needed within the said system. These queries validated the results of service discovery as well as sub assumption conditions within the ontology.

Conclusion:

Functional Group Ontology (FGO) proves to an effective, easy and explicit knowledge representation for chemical functional group for annotating chemical components. The relevance of the FGO goes beyond annotation and information retrieval applications. It has been shown that FGO may facilitate large-scale predictive applications in functional moieties and classification responsible for biological activity. We have also used the semantic similarity score to cluster the small molecules which is much faster than graph based similarity measure but substantially extracted, mined and accommodated pharmacophoric pattern in similar cluster. Further chemical structures sharing similar annotation can be analyzed to extract out the common pattern and the coordinate information content of the FGO is helpful for finding pharmacophore pattern. The content based searching through the semantic similarity and annotation is also helpful for sub structure searching in ligand database and also in information content retrieval from classified databases of ligands. This can be also helpful in fragment databases used in combinatorial designing and de-novo ligand growing.

Acknowledgment:

We gratefully acknowledge partial support we received for this work in the form of grant-in-aid received from DST,

Govt. of India, for the centre-project Indo-Russian Centre of Biotechnology at Indian Institute of Information Technology (IIIT), Allahabad, India. We would also like to acknowledge the helpful discussions and constructive suggestion we have with Dr. Sudip Sanyal, IIIT, Allahabad, India.

References:

- [01] W. Christopher, *Ontology Research*, 24: 11 (2003)
- [02] N. Tilford, *Proceeding of Sixth Annual Bio-Ontologies Meeting*, 31 (2003)
- [03] P. G. Baker, *et al.*, *Bioinformatics*, 15: 510 (1999) [PMID: 10383475]
- [04] Z. B. Miled, *et al.*, *Online journal Bioinformatics*, 1: 60 (2002)
- [05] M. Ashburner, *et al.*, *Nat. Genet.*, 25: 25 (2000) [PMID: 10802651]
- [06] C. Brooksbank, *et al.*, *Nucleic Acids Res.*, 33: D46 (2005) [PMID: 15608238]
- [07] H. J. Feldmann, *et al.*, *FEBS Lett.*, 579: 4685 (2005) [PMID: 16098521]
- [08] <http://merian.pch.univie.ac.at/~nhaider/cheminf/cm mm.html>
- [09] P. Resnik, *J. Art. Int. Res.*, 11: 95 (1999)
- [10] H. M. Berman, *et al.*, *Nucleic Acids Research*, 28: 235 (2000) [PMID: 10592235]
- [11] W. H. Clyd & K. D. Joshi, *Communications of the ACM*, 45: 42 (2002)
- [12] R. McEntire, *et al.*, *ISMB*, 239 (2000)
- [13] www.mdli.com
- [14] D. Weininger, *J. Chem. Inf. Comput. Sci.*, 28: 31 (1988)

Edited by R. Sowdhamini

Citation: Varadwaj & Lahiri, *Bioinformatics* 2(3): 113-118 (2007)




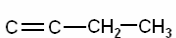


License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Supplementary material

Definition of equation 1

$$\text{SIM}(L_i, L_j) = \frac{1}{m+n} \times \sum \text{sim}(s_i, s_j) \quad \rightarrow \quad (1)$$

Few annotation examples:

IUPAC name	Chemical Formulae	Structure	FGO annotation
Ethanoic acid	CH ₃ COOH		C1C(1O)2O
Ethanol	C ₂ H ₅ OH		C1C1O
Propanone	CH ₃ COCH ₃		C1C(1C)2O
But-1-ene	C ₄ H ₈	 C=C-CH ₂ -CH ₃	C2C1C1C
Prop-1-yne	C ₃ H ₄	 ≡-CH ₃	C3C1C
Benzene	C ₆ H ₆		{c1c2c1c2c1c}

Query functional group	No. of Hits with semantic score > 0.3
Carbonyl group	6276
Carboxylic group	2973
Acetal group	867
Amide group	1961
Ester group	135

Table 1: Hit found with different functional group using semantic similarity index (SIM > 0.3)