Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Functional group based ligand binding affinity scoring function at atomic environmental level

Pritish Kumar Varadwaj[1,] * and Tapobrata Lahiri[1]

[1]Indian Institute of Information Technology, Allahabad, 211012, India;
Pritish Kumar Varadwaj* - Email: pritish@iiita.ac.in; * Corresponding author

received September 28, 2008; accepted December 06, 2008; published January 12, 2009

Abstract:
Use of knowledge based scoring function (KBSF) for virtual screening and molecular docking has become an established method for drug discovery. Lack of a precise and reliable free energy function that describes several interactions including water-mediated atomic interaction between amino-acid residues and ligand makes distance based statistical measure as the only alternative. Till now all the distance based scoring functions in KBSF arena use atom singularity concept, which neglects the environmental effect of the atom under consideration. We have developed a novel knowledge-based statistical energy function for protein-ligand complexes which takes atomic environment in to account hence functional group as a singular entity. The proposed knowledge based scoring function is fast, simple to construct, easy to use and moreover it tackle the existing problem of handling molecular orientation in active site pocket. We have designed and used Functional group based Ligand retrieval (FBLR) system which can identify and detect the orientation of functional groups in ligand. This decoy searching was used to build the above KBSF to quantify the activity and affinity of high resolution protein-ligand complexes. We have proposed the probable use of these decoys in molecular build-up as a de-novo drug designing approach. We have also discussed the possible use of the said KSBF in pharmacophore fragment detection and pseudo center based fragment alignment procedure.

Keywords: ontology; small molecule; functional group; pharmacophore; semantic similarity; scoring function; knowledgebase; Distance based scoring

Background:

The number of publicly available protein structures in the Research Collaboratory for Structural Biology database has grown to more than 30,000 structures with thousands of them being added each year. In addition the number of structures of small molecules available in public databases and proprietary databases has reached into the millions. This wealth of available data raises the question of how it can be best used to assist in drug design and discovery. Moreover the process of novel lead finding for a new drug target has became the most important and undoubtedly one of the most crucial steps in a drug development program. These days researchers are following mostly two complementary strategies 1) experimental high-throughput screening to discover possible leads from large compound libraries, and 2) computational methods exploiting structural information of the protein binding site to discover new lead by virtual screening of large databases **[1, 2, 3, 4]**. The virtual screening approaches try to predict the actual binding mode of a ligand at the binding site by scoring each possible binding mode through docking. The in silico virtual screenings are useful because they are fast enough to scan over several hundred to thousand compounds **[5, 6]**. Through virtual screening we can rank the possible modes of ligand binding and also can predict the Gibbs free energy of binding, provided that the structural information of receptor is known and the scoring function is good enough to do so. Usually the performance of such methods is determined by assessing whether the binding geometry of protein-ligand complexes resolved by X-ray crystallography or NMR is reproduced. This validation criterion imposes some preconditions onto the methods being developed because, of the availability of limited resolution protein-ligand complexes.

There are two broad categories of scoring functions. The first category of functions are largely based on some aspects of the known physics of molecular interaction, such as the van-der Waals force, electrostatics, and the bending and torsional forces, to determine the energy of a particular conformation **[7-12]**. The second categories of functions are knowledge-based. Each of these knowledge-based functions tries to capture some aspects of the protein- ligand complex native conformations, such as the tendency of a certain amino acid to be exposed or buried relative to the solvent and its distance from interacting groups of ligand. These knowledge-based functions are compiled based on the statistics of a database of experimentally determined protein-ligand complex structures **[13-23]**.

Interaction between these two categories of functions has resulted in a fertile ground for the experimentation and construction of new scoring functions. The distance based scoring function existing in today's paradigm take atom as a single moiety and hence missing the environmental effect of it. We formulate and analyze an analogous knowledge-based scoring function which involves the distance of functional group from triplets of residues in a protein conformation. The functional group scoring took in to account the environmental effect of atoms and hence considering functional group singularity concept unlike atom singularity discussed in above methods. We also investigate the effect of using various approaches for compiling the prior distribution on the performance of the knowledge-based function.

We first briefly review the existing knowledge based scoring function approaches. We then describe the construction of a knowledge-based scoring function which incorporates environmental effect of atom under consideration. The performance of the proposed knowledge based function in protein-ligand binding affinity study. Finally, we propose some possible extensions to the current form of the scoring function.

Theoretical background and Methodology adopted:

Existing methodologies

Knowledge-based potentials have been applied successfully to rank different solutions of the protein-folding problem **[24-26]**. This approach has also been applied to several case studies for the ranking of different protein-ligand complexes. None of these, however, engaged in identifying environmental effect of ligand atoms from which the distances were calculated to residues triplet of protein. Wallqvist and co-workers **[27, 28]** classified the surfaces of buried ligand atoms found in 38 complexes to develop a model for Gibbs free energy prediction of binding based on these observed atom-atom preferences. Analyzing ten HIV protease inhibitor complexes, they approximated the free energy of binding to an accuracy of +/- 1.5 kcal/mol. Verkhivker and coworker **[29]** using a data set of 30 HIV-1, HIV-2, and SIV proteases, compiled a distance-dependent knowledge-based pair-potential which was then combined with the hydrophobicity **[30]** and conformational entropy scales **[31]** that originally had been developed to explain protein folding and stability. Muegge and Martin **[32]** explored structural information of known protein-ligand complexes from the PDB and derived distance-dependent Helmholtz free interaction energies of protein ligand atom pairs. Tested on 77 protein-ligand complexes, the calculated score displayed a standard deviation from the observed binding affinities of 1.8 log Ki units. The scoring function was further evaluated by docking weak-binding ligands to the FK506-binding protein **[33]**. Similarly DeWitte and Shaknovich **[34]** used a sample of 126 structures from the PDB to develop a set of inter-atomic interaction free energies for a variety of atom types. Gohlke et. al. **[35]** has studied 91 protein-ligand PDB complexes to calculate a new scoring function to discriminate and predict ligand-binding modes in multiple solutions.

Limitation of existing methodologies and our approach

All of the existing approaches discussed above calculate the distances between ligand atoms in consideration to the functional moieties within target active site. The functional moieties of active site can be the whole residue triplet or only an individual atom of interacting amino acids (Figure 1). In all cases distances were taken and calculated over series of high resolution Protein-ligand complexes to come up with some consensus distance score, each with a measured standard deviation.

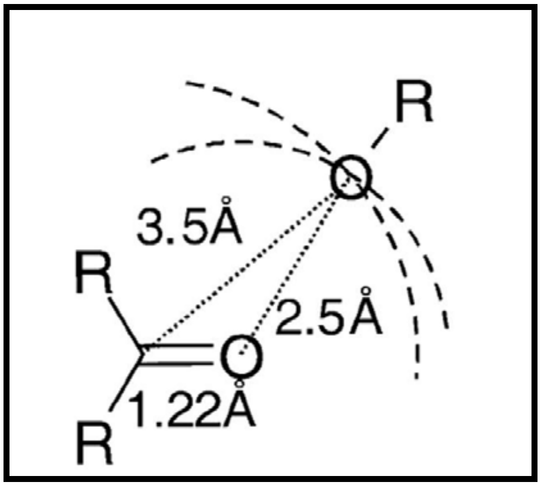


Figure 1: Atom Singularity concept: each atom treated as individual entity irrespective of the functional groups.

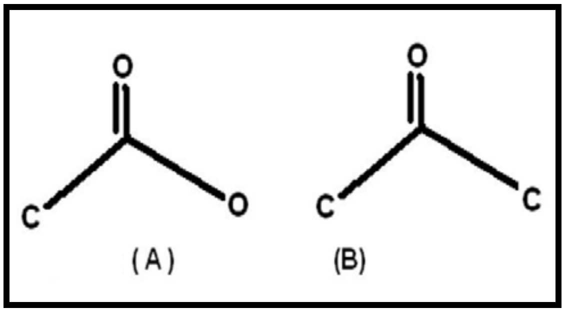


Figure 2: Oxygen in different environment, (A) carboxylic acid and (B) ketone.

But one basic problem associated with these above set of discussed scoring function is tough it has statistical account for all distances measure, it neglects the atomic environment. From Figure-2 it is self evidenced that the Oxygen in Acid (A) and Ketone (B) group will have different electronic cloud distributions associated with it. So it should be treated separately unlike the atom singularity concept (single distance measure) in earlier distance based scoring functions. We have taken this in to account in our study to design a scoring function which is functional group based. Hence atomic environmental effects have been considered, in our functional group singularity concept.

Identification of functional group

We have designed a tool which is able to automatically detect and assign functional group (FG) information to any given small molecule **[36]**. Theoretically it is able to identify any functional group though only 210 different functional groups have been annotated from our ligand dataset. 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.

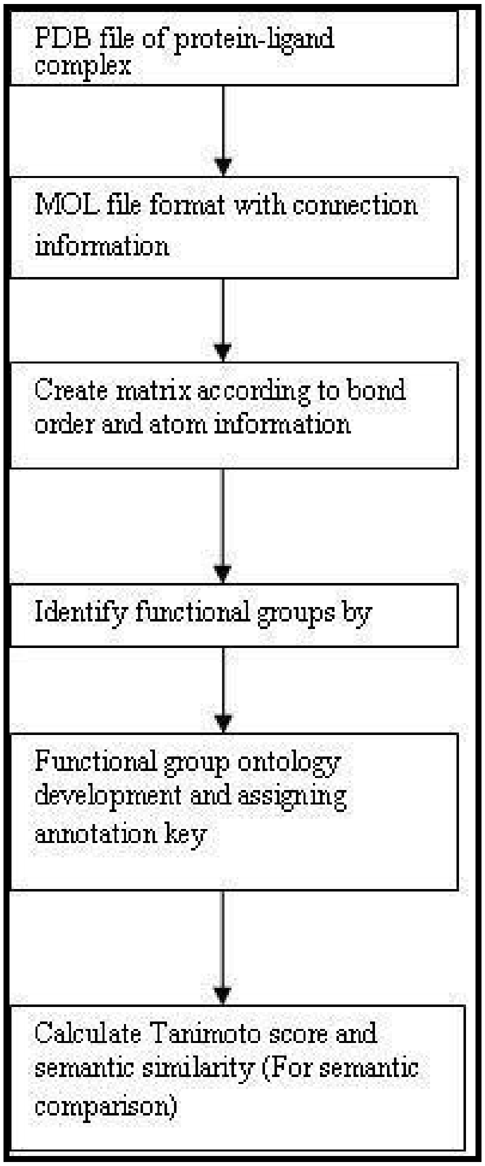

PDB file of protein-ligand complex
MOL file format with connection information
Create matrix according to bond order and atom information
Identify functional groups by
Functional group ontology development and assigning annotation key
Calculate Tanimoto score and semantic similarity (For semantic comparison)

Figure 3: Flow chart of FGO generation.

Ligand files preparation

With an aim to make our functional group study more meaningful and ontology designing more application oriented, we have specifically taken selected set of protein-ligand complexes with high resolution (more than 2 angstrom) from Protein Data Bank (PDB) as the input set for our study. The input set of PDB file has coordinate information about the protein and ligand bound to it. These PDB file have no information about the atomic connections hence, we have converted the files to MOL format. MOL format file has atomic as well as connection information with all specific relevant information of PDB file restored. We have specifically extracted the heteroatom information of PDB file with row id. HETATM and its connection information from MOL file.

Coordinate information of Water molecules are explicitly removed and only small molecule coordinates information were taken. Adjacency matrix was developed from the above flat file information and further processed for functional group identification. The highest bond order priority based link finding was done from the matrix data. Corresponding atom information with each bond order was carried out to categorize each probable functional group in given ligand. Information was mapped to chemical tree for functional group identification and annotation in descending order of bond order. Annotation key were developed and used in FGO to identify and represent the functional group along with coordinate and other information.

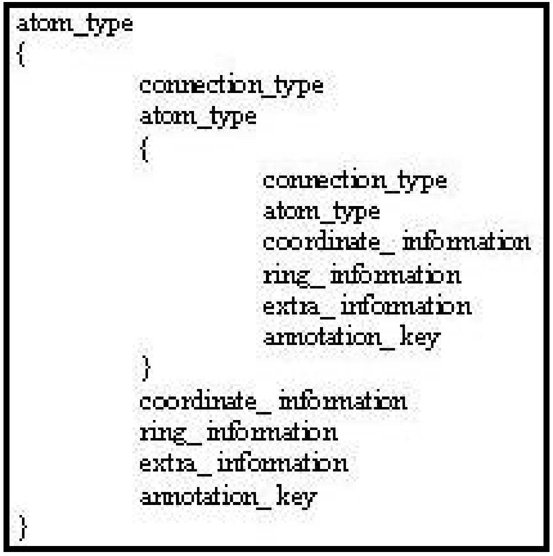
Functional group ontology

The Functional group ontology (FGO) designed to treat each functional group as an abstract super class with each atom of it as an object **[36]**. The detail information of atom like and its connection information with parent atoms were stored in a single class. Such classes were hierarchically arranged to denote each functional group. Refer supplementary material for parameter definitions.

Discussion:

In total 250 proteins structure has been studied to estimate *ΔP* for each set of available functional groups. The distance between CG of functional group and interacting functional point of active site triad were measured with in cutoff of 12 Å. Here we have made a comparative study between oxygen of carbonyl group in ketones and carboxylic acids functional groups. Out of selected 250 high resolution structures only 169 was to found to have one or more of the above mentioned functional groups. Distances of calculated radii for statistical potential function for these groups were calculated and frequency distribution plot for both functional groups has been shown in Figure 5 and 6 respectively.

From above Figure 5-6 it's clearly evidenced that though both the functional group in study contains carbonyl moiety explicitly, but due to environmental difference in both cases the mean distance 'r' differs in both case. The carboxylic acid is having distance radii between 2.75 Å to 2.95 Å where as ketone group is falling between 2.35 Å to 2.64 Å. It suggests a strong protein-ligand interaction prevails in with ketone group in comparison to acid group in protein active site pocket. It can be further evidenced as the carboxylic acid group is having one oxygen atom in neighborhood of carbonyl group hence the strong electonegetivity of oxygen atom make the electronic cloud over carbonyl group less dense to form a weak interaction with responsible amino acid residue in active site pocket. Whereas absence of electronegative atom in neighborhood of ketone functional group makes it to interact with active site stronger than that of carboxylate.



```
atom_type
{
        connection_type
        atom_type
        {
                connection_type
                atom_type
                coordinate_ information
                ring_ information
                extra_ information
                annotation_ key
        }
        coordinate_ information
        ring_ information
        extra_ information
        annotation_ key
}
```

Figure 4: Class architecture of single atom used in FGO.

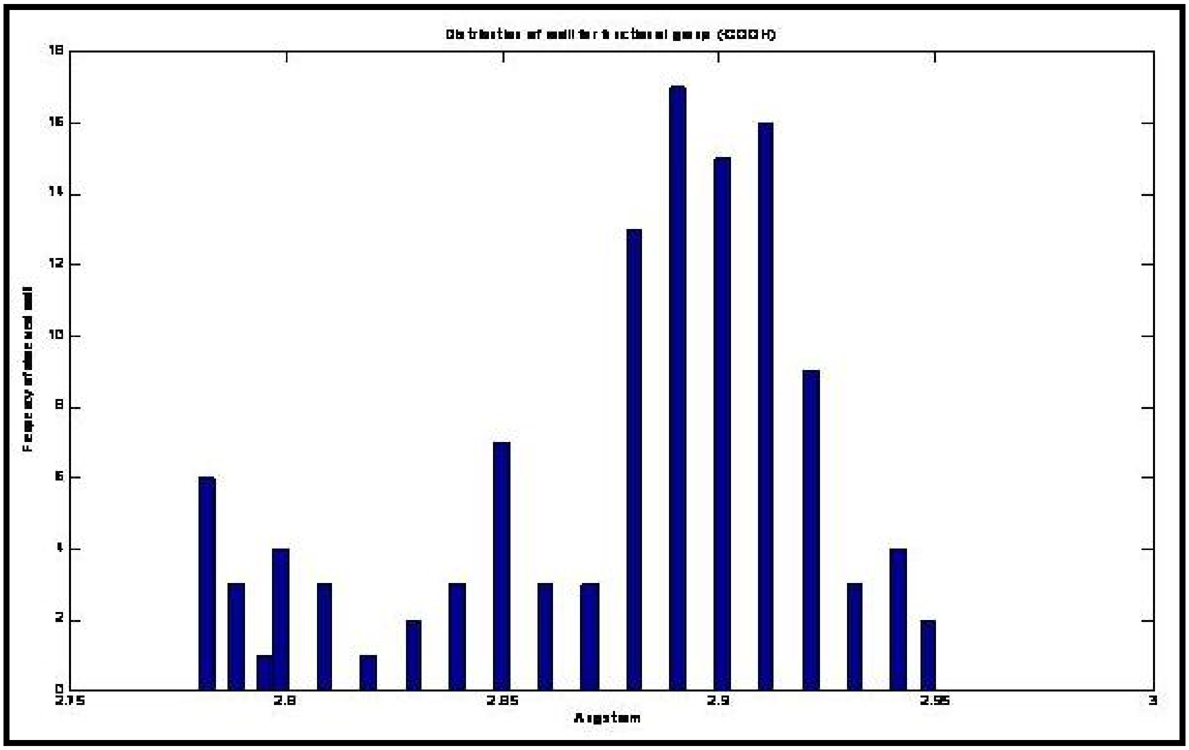


Figure 5: Distance frequency distribution for ketone functional group in protein active site.

This work has an account of responsible functional group fragment hence can be used further in de-novo drug designing, where molecular buildup approach can be used to generate the new lead for a receptor. Orientation profile of all functional group is also represented in the ontology said hence can be used to explicitly use the orientation placement of these blocks in receptor cavity to build the final lead in de-novo build-up approach. Semantic similarity discussed above can be used for pharmacophore matching and database searching algorithm and also it can be used for structural pharmacophore clustering of ligand molecules. The diversity management in virtual screening can be also addressed by using the semantic similarity score.

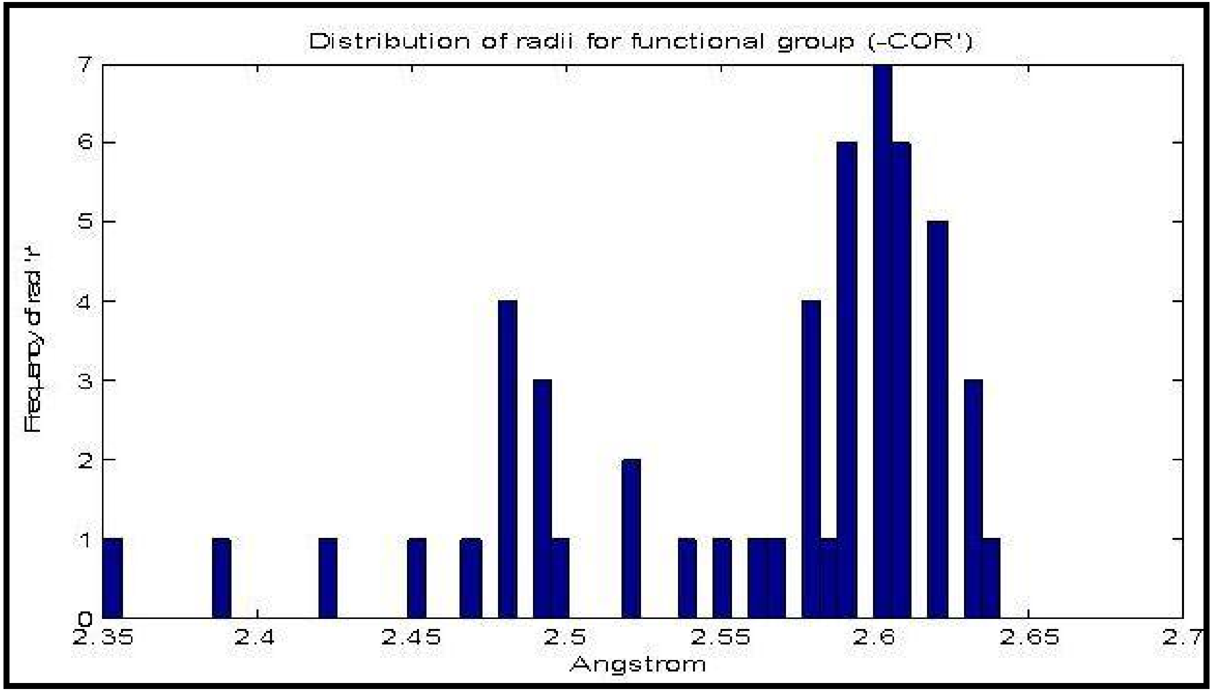


Figure 6: Distance frequency distribution for carboxylic functional group in protein active site.

Conclusion:

In this study, we have constructed and analyzed a functional group based distance knowledge-based scoring function. The scoring function is inspired by the previous work of Gohlke *et. al*. **[35]**, who designed a atomic distance based scoring function new scoring function to discriminate and predict ligand-binding modes in multiple solutions. Our formulation of the functional group ligand retrieval system can retrieve ligand functional group fragments and can represent it by uniquely designed ontology. This ontology has also exact coordinate information to track the specific orientation of ligand binding mode. We have used the ligand functional group to create a pseudo centre distance based scoring function which can used as distance-dependent potential function of Protein-ligand binding. In our numerical distance approximation experiments we have used an r-min of 1 Å and r-max of 12 Å to accommodate bulky functional group orientation to produce good results. It took atomic environment in to consideration and hence more accurately reveal and incorporate the ligand electronic cloud distribution in to scoring function. We have taken a comparative assessment between carbonyl atom of ketone and acid and the result shown above support the fact that, the environmental effects make the two set of carbonyl atom quite distinct from each other. This statistical distance potential can be used for measuring ligand-protein binding affinity.

References:

[01] H. Kubinyi, *Curr. Opin. Drug Discov. Devel.,* 1: 4 (1998)
[02] K. Muller, *Perspectives in Drug Discovery and Design,* 3: 234.1 (1995)
[03] J. H. Van Drie and M. S. Lajiness, *Drug Discov. Today*, 3: 274 (1998)
[04] W. P. Walters *et al., Drug Discov. Today,* 3: 160 (1998)
[05] I. D. Kuntz *et al., J. Mol. Biol*., 161: 269 (1982)
[06] T. Lengauer and M. Rarey, *Curr. Opin.Struct. Biol.,*.6: 402 (1996)
[07] B. Brooks *et al*., *J. Comput. Chem.,* 4: 187 (1983)
[08] W. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.,* 110: 1657 (1988)
[09] G. Nemethy *et al., J. Phys. Chem*., 96: 6472 (1992)
[10] W. D. Cornell *et al., J. Am. Chem. Soc.,* 117: 5179 (1995)
[11] S. Weiner *et al., J. Comput. Chem*., 7: 230 (1986)
[12] A. D. MacKerell *et al. J. Phys. Chem.,* 102: 3586 (1998)
[13] S. Wodak and M. Rooman, *Curr. Opin. Struct. Biol.,* 3: 247 (1993)
[14] D. Gilis and M. Rooman, *J. Mol. Biol.,* 257: 1112 (1996)
[15] J. Moult *et al., Proteins,* 29: 2 (1997)
[16] J. Moult *et al., Proteins,* 37: 2 (1999)
[17] J. Moult *et al., Proteins,* 45: 2 (2001)
[18] J. Moult *et al., Proteins,* 53: 334 (2003)
[19] R. Samudrala and M. Levitt, *BMC Struct. Biol.,* 2: 3 (2002)
[20] R. Samudrala and J. Moult, *J. Mol. Biol.,* 275: 895 (1998)
[21] R. Samudrala *et al., Proteins,* 3: 194 (1999)
[22] K. Dunker *et al.*, *Proceedings of the Pacific Symposium on Bio-computing, World Scientific Press, Singapore,* 505.
[23] L. M. Sipp, *Curr. Opin. Struct. Biol.,* 5: 229 (1995)

[24] R. L. Jernigan and I. Bahar, *Curr. Opin. Struct. Biol.,* 6: 195 (1996)

[25] S. Vajda e*t al., Curr. Opin. Struct. Biol.,* 7: 222 (1997)

[26] A. E. Torda, *Curr. Opin. Struct. Biol.,* 7: 200 (1997)

[27] A. Wallqvist and D. G. Covell, *Proteins: Struct. Funct. Genet.,* 25: 403 (1996)

[28] A. Wallqvist *et al., Protein Sci.,* 4: 1881(1995)

[29] G. Verkhivker *et al., Protein Eng.,* 8: 677 (1995)

[30] K. A. Sharp *et al., Biochemistry,* 30: 9686 (1991)

[31] S. D. Pickett and M. J. Sternberg, *J. Mol. Biol.,* 231: 825 (1993)

[32] I. Muegge *et al., J. Med. Chem.,* 42: 2498 (1999)

[33] I. Muegge and Y. C. Martin, *J. Med. Chem.,* 42: 791 (1999)

[34] R. S. DeWitte and E. I. Shaknovich, J. *Am. Chem. Soc.,* 118: 11733 (1996)

[35] H. Gohlke, *et al., J. Mol. Biol.,* 295: 337 (2000)

[36] V. P. Kumar *et al., Bioinformation,* 2: 113 (2007)

[37] H. M. Berman *et al., Nucleic Acids Research,* 28: 235 (2000)

Edited by P. Kangueane
Citation: Varadwaj & Lahiri, Bioinformation 3(6): 268-274 (2009)

Supplementary material

Parameter definitions

Semantic similarity:

Let S be the set of terms used in the FGO (functional group ontology). Information-theoretic approaches to measuring similarity between terms, $\boxed{s \in S}$, 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

$$\boxed{\mathrm{SIM}(\mathrm{L_i},\mathrm{L_j}) = \frac{1}{\mathrm{m}\times\mathrm{n}} \times \sum \mathrm{sim}(\mathrm{s_i},\mathrm{s_j})}$$

Protein-Ligand complex:

We have selected 250 protein PDB file from Research Collaboratory for Structural Biology database **[37]** (RCSB, www.rcsb.org). The above set of high resolution (2 Å or less) structure of Protein – ligand complex i.e. Protein with its ligand were taken and preprocessed for scoring function calculation.

1. The coordinate information of functional groups detected by FGO was taken to calculate central CG point. This act as a pseudo point for distance calculation.

$$\boxed{CG(FG) = \frac{1}{N} \times \sum_{i=1toN} (Xi + Yi + Zi)}$$

2. The distances between the CG and active site triads were calculated for all protein-ligand complexes.
3. Standard deviation of each of such functional group distance distributions were calculated over range of protein-ligand complex.

Protein-Ligand scoring function

Given a protein *I* and a ligand *J*, interactions between protein and ligand atoms to a pseudo potential $\Delta P_{I,J}$ by means of statistically derived distance-dependent pair potentials $\Delta P_{i,j}\,(r)$:

$$\boxed{\Delta P_{i,j} = \sum_{i \in I} \sum_{j \in J} \Delta P_{i,j}(r)}$$

These pair potentials *$\Delta P_{i,j}(r)$* between *i* (atom of functional triad in protein cavity) and j(CG of functional group in ligand), located at a distance *r*, can be obtained from the normalized radial pair distribution function $f_{i,j}(r)$ and the normalized mean radial pair distribution function $f(r)$.

$$\boxed{\Delta P_{i,j}(r) = -\ln\frac{f_{i,j}(r)}{f(r)}; f(r) = \frac{\sum_{i}\sum_{j} f_{i,j}(r)}{i*j}}$$

The normalized radial pair distribution function $f_{i,j}(r)$ is compiled from occurrence frequencies *N* of atom pairs with types *i* and *j* in a molecular database

$$\boxed{f_{i,j}(r) = \frac{N_{i,j}(r)/4\Pi r^2}{\Sigma(N_{i,j}(r)/4\Pi r^2)}}$$

Where distances $1.0 \leq r \leq 12$ Å are considered.
Scaling to $4\pi r^2$ accounts for the volume of the spherical shell of the radius ‘r’ and the thickness dr. The number Ni,j(r) of distance pairs *i,j* at a distance between *r* and *r + dr* is obtained by counting the occurrences:

$$\boxed{N_{i,j}(r) = \sum_{i}\sum_{j}\delta(|\vec{r}_i - \vec{r}_j|, r)}$$

Where the double summation runs over all atom / pseudo atom types i and j present in the database, respectively. The minimal distance boundary *rmin* was set to 1 Å and the maximal distance bound *rmax* was set to 12 Å.