Generating "Fragment-based Virtual Library" using Pocket Similarity Search of Ligand-Receptor Complexes

Raed S. Khashan

Running head:

Virtual Library of Fragments

Raed S. Khashan, PhD

Assistant Professor

Department of Pharmaceutical Sciences

College of Clinical Pharmacy

King Faisal University

Al-Ahsa, KSA 31982

E-mail: rkhashan@kfu.edu.sa

Summary

As the number of available ligand-receptor complexes is increasing, researchers are becoming more dedicated to mine these complexes to aid in the drug design and development process. We present free software which is developed as a tool for performing similarity search across ligand-receptor complexes for identifying binding pockets which are similar to that of a target receptor. The search is based on 3D-geometric and chemical similarity of the atoms forming the binding pocket. For each match identified, the ligand's fragment(s) corresponding to that binding pocket are extracted, thus, forming a virtual library of fragments (FragVLib) that is useful for structure-based drug design. The program provides a very useful tool to explore available databases.

Key words: Fragment-based, drug design, virtual library, in-silico, pocket similarity, and sub-graph mining.

1. Introduction

We present a tool that mine ligand-receptor complexes and generate a library of fragments for a target receptor so it can be used for structure-based drug design, such as Fragment-Based Lead Design (FBLD). FBLD is a computational approach which begins with a small low affinity fragment(s) which bind to the target of interest, followed by a careful construction and optimization of these fragments to end up with a high affinity lead drug. In theory, this is a highly efficient approach for drug design, and it has become enormously popular in the past few years ***(1-4)***.

Our method, FragVLib ***(5)***, relies on a *Graph*-based representation for interfacial atoms of a ligand-receptor complex. Interfacial atoms are defined as the adjacent receptor and ligand atoms which are within certain cutoff distance. Interfacial atoms are represented by nodes, and distances between them are represented by edges connecting these nodes. Therefore, the resulting interfacial-graph contains a number of nodes representing atoms from the ligand connected by edges to a number of nodes representing atoms from the receptor. Furthermore, the interfacial-graph also includes all the atoms that are covalently bound to the interfacial atoms. These atoms are represented by nodes, and the covalent bonds connecting them to the interfacial atoms are represented by edges (**Fig. 1**).

We should mention that we make use of the tessellation technique to identify the interfacial atoms. Specifically, we use Almost-Delaunay (AD) tessellation ***(6)*** which has a unique advantage of incorporating the imprecision of the point coordinates in defining the tessellation patterns. Besides the cutoff distance (*ADdistance*) used to identify adjacent atoms, a threshold value (*ADepsilon*) is used to signify the minimum perturbation needed for an atom to be part of the interfacial atoms. This is important when dealing with bad resolution ligand-receptor complexes.

Now let's assume that we have a "target" ligand-receptor complex for which we are interested in designing a lead compound using FragVLib method. Let's also assume that we have a database of X-ray crystallized ligand-receptor complexes, i.e., "native"

complexes. First, we will generate the interfacial-graphs for all ligand-receptor complexes involved, i.e., the target complex and all the native complexes.

Now since we have the complexes' interfaces represented by interfacial-graphs, we can use efficient sub-graph match to perform a pocket similarity search between the interfacial-graph of the target complex and the interfacial-graph of each one of the 'native' complexes in the database. The match considers all possible sub-graphs and is performed over the atoms and bonds composing the receptor side of the interfacial graphs only; this is a pocket similarity search, and ligands were only used to define the binding pockets. The match takes into account the chemistry and the 3D geometry of the matching atoms and bonds. The 3D geometry is checked by making sure that the matching atoms super-impose within a user defined RMSD cutoff value (*dRMSDcutoff*). The user of the tool (FragVLib) can also limit the size of an accepted match (i.e., number of nodes in the matched sub-graph) by providing the minimum value (*minMatchSize*) and a maximum value (*maxMatchSize*) for the size.

Every time an accepted sub-graph match is identified between the interfacial-graph of the target complex and the interfacial-graph of a native complex, the ligand's part (atoms and bonds) of the interfacial-graph of the native complex that are only in direct contact with the identified sub-graph match are copied into the pocket of the target receptor. When repeating this pocket similarity search using each native complex in the database, we will generate a collection of chemical fragments filling the binding pocket of the target

receptor. These fragments constitute the so called ‘Fragment-based Virtual Library’ or FragVLib (**Fig. 2**).

Finally, for lead design, the user can explore these fragments and perform one of the following: Growing them into the depth of the binding pocket; carefully connect two or more fragments into one compound for optimized potency; or, merge two or more fragments in regions of mutual overlap to construct a lead compound ***(7)***.

2. Materials

The program is written in C++, and it is publicly available freeware; it can be copied and distributed freely. The user manual and the pre-compiled executables can be downloaded by going to the website “http://www.bioinformatics.org/fragvlib” and installing the file “FragVLib.zip”. It is easy to install (no external libraries) and easy to use as we will explain in the next section. After unzipping the file, you will have the following executables:

- *getIntGraph4Target*
- *getIntGraphs4DB*
- *FragVLib*
- *rmLigHs*
- *rmProHs*
- *rmProWs*
- *getAlmDisGraphMol2*
- *mol2graphXYZ*

- *ADCGAL*

- *ADedgeCGAL*

Notice that all of them run on a Linux operating system. You will have the target receptor-ligand complex for which you would like to design the lead compound in the PDB file format and in MOL2 file format, for the receptor and the ligand, respectively. You will also have a database of native, X-ray crystallized, receptor-ligand complexes in the same file format. You will need a program like *PyMOL* to view the fragments after generating them; you can install it from this website: "http://www.pymol.org/".

3. Methods

The following are steps you will need to generate the virtual library of fragments (FragVLib). You need to have all the executable files and your files in one directory.

3.1 Obtaining the interfacial-graph for the target receptor

The first step in this method is obtaining the interfacial-graph for the target receptor. You should have the target receptor-ligand complex available in MOL2 file format for the ligand, and in PDB file format for the receptor. Then you can type the following command:

getIntGraph4Target namesFile ADdistance ADepsilon noW

The *namesFile* is a file containing the name (including location) of the ligand's file, followed by space, followed by name (including location) of the receptor's file. The *ADdistance* is the maximum distance for two interfacial atoms to be considered in

contact, the recommended value is between 3.5 to 5.8 Å. The *ADepsilon* parameter is the maximum perturbation allowed for the location of an atom, the recommended value is between 0.01 to 0.1 Å. Go back to the 'Introduction' section for more details about these parameters (*also, see* **Note 2**). The *noW* is a parameter that, if included, tells the program to ignore water molecules and treat them implicitly (*see* **Note 3**). If you want water molecules to be part of the interface, simply do not include this parameter. Below are two examples of running the *getIntGraph4Target* program:

getIntGraph4Target namesFile 4.25 0.05 noW

getIntGraph4Target namesFile 4.0 0.01

The output of this step will be two files for the atoms and bonds of the target receptor's interfacial-graph: *Target_atomsXYZ*, and *Target_bonds*.

3.2 Obtaining the interfacial-graphs for the database of complexes

The second step is obtaining the interfacial-graphs for the database of X-ray crystallized (native) receptor-ligand complexes. For each complex, you should have the ligand's file in MOL2 file format, and the receptor's file in PDB file format. You need to list the names of all receptor-ligand complexes in one file *namesFile*, such that each line refers to one complex and contains the name (including location) of the ligand's file, followed by space, and followed by the name (including location) of receptor's file. Then you will type the following command:

getIntGraphs4DB namesFile ADdistance ADepsilon noW

Make sure you use the same values for parameters *ADdistance* and *ADepsilon* used in previous step when obtaining interfacial-graph for the target receptor. The output of this step will be two files for the atoms and bonds of the interfacial-graphs: *DB_atomsXYZ_name*, and *DB_bonds*.

3.3 Generating the virtual library of fragments

Finally, the last step is performing the pocket similarity search between the target receptor's interfacial-graph, and the interfacial-graph for each receptor-ligand complex in the database. A subgraph match will start by running the following command:

FragVLib Target_atomsXYZ Target_bonds DB_atomsXYZ_name DB_bonds minMatchSize minMatchSize dRMSDcutOff outDir

The first four files are the same ones generated in the previous two steps, so you will not have to do anything about them. The *minMatchSize*, and *maxMatchSize* is the minimum and maximum size of a matched interface to be accepted (*see* **Note 4**). The *dRMSDcutoff*, is the maximum value for an RMSD of the matching pockets to be accepted as similar pockets, it can takes value from 0.1 to 1.0 Å. Go back to the 'Introduction' section for more details about these three parameters (*also, see* **Note 5**). The *outDir* is the directory where all the generated fragments will be stored in (*see* **Note 6**). These fragments will constitute the virtual library, and they will be stored in MOL2 file format. You can use *PyMOL* to view the fragments and start the lead design process.

4. Notes

1. The program utilizes efficient tools for representing the interfacial atoms of the ligand-receptor complexes, as well as performing the pocket similarity search. However, the major drawback for the method is the fact that it relies on sub-graph matching as a way of performing the match searching process. Sub-graph mining in the presence of isomorphism is a well known NP-Complete problem ***(8)*** in the field of computer science. Such kind of problems is typically solved using techniques such as: Approximation, Randomization, Parameterization, Restriction, and Heuristic algorithms. Herein, to speed up the searching process, we implemented parameterization, restriction and heuristic algorithms. Parameterization is possible by controlling certain input parameters, such as: *ADdistance*, *ADepsilon*, *minMatchSize*, *maxMatchSize*, and *dRMSDcutoff*. For example, using short cutoff distances (*ADdistance*) in identifying interfacial atoms will result in interfacial-graphs that are smaller in size, and therefore, faster search is obtained.
2. Short cutoff distances (*ADdistance*) can be used when the target receptor's binding pocket is expected to have interactions such as: hydrogen-bond, and ion exchange, which occur over short distances. If we expect hydrophobic interactions, which can occur over large distances, higher cutoff values can be used.
3. Water molecules can be included as part of the interface, or they can be omitted and treated implicitly by adding the *noW* parameter. Omitting water molecules will speed up the search process.

4. You can modify the size of the matching binding pockets to search for a smaller binding region in the target receptor by modifying values of *minMatchSize*, and *maxMatchSize*.
5. The RMSD cutoff value (*dRMSDcutoff*) for accepting the matched (super-imposed) interfacial-graphs can be used to decide on how (geometrically) similar the binding matching binding pockets are.
6. If you decide to run another round of *FragVLib*, make sure you choose a different name for the *outDir*, or delete the one you have.
7. Always make sure you have all the executables (listed in 'Materials' section) in the same directory where you are running the program.

5. References

1. Hajduk, P.J. and Greer, J. (2007) A decade of fragment-based drug design: strategic advances and lessons learned. *Nat Rev Drug Discov* 6, 211-219.

2. Loving, K., Alberts, I., and Sherman, W. (2010) Computational approaches for fragment-based and de novo design. *Curr Top Med Chem* 10, 14-32.

3. Rees, D.C., Congreve, M., Murray, C.W., and Carr, R. (2004) Fragment-based lead discovery. *Nat Rev Drug Discov* 3, 660-672.

4. Schneider, G. and Fechner, U. (2005) Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* 4, 649-663.

5. Khashan, R. (2012) FragVLib a free database mining software for generating "Fragment-based Virtual Library" using pocket similarity search of ligand-receptor complexes. *J Cheminform* 4, 18.

6. Bandyopadhyay, D. and Snoeyink, J. (2007) Almost-Delaunay simplices: Robust neighbor relations for imprecise 3D points using CGAL. *Computational Geometry* 38, 4-15.

7. Detering, C., Gastreich, M., and Lemmen, C. (2010) Leading fragments to lead structures: Fragment evolution, merging and core replacement, and docking. *Chemical Information Bulletin* 62, 62.

8. Huan, J., Wang, W., and Prins, J. (2003) Efficient Mining of Frequent Subgraph in the Presence of Isomorphism. *In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM): 19-22 November 2003, Melbourne, Florida.*

Figure Captions

Fig. 1. (a) An example of a receptor-ligand complex. (b) The same example after defining interfacial atoms using Almost Delaunay (AD) tessellation. (c) The interfacial atoms and their bonds form the interfacial-graph.

Fig. 2. A picture for the target receptor-ligand complex on the left side, and another picture for the receptor after identifying the fragments (virtual library of fragments) using FragVLib, on the right side.