

# A New 3D Statistical Potential to Predict Protein-Ligand Interactions Based on Atomic Interaction Patterns in PDB

Ernesto Moreno<sup>1</sup>, Luis A. Diago<sup>2</sup>

<sup>1</sup> Centro de Inmunología Molecular, P.O.Box 16040, Havana 11600, Cuba  
emoreno@ict.cim.sld.cu

<sup>2</sup> Instituto Superior Politécnico, Havana  
diago@electronica.cujae.edu.cu

**Keywords.** Protein-ligand interaction, Docking, Knowledge-based potential, Protein-ligand database

## 1. Introduction

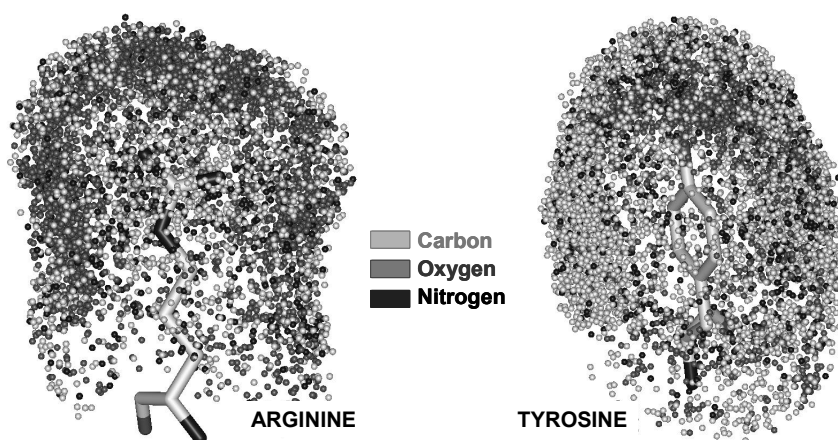
Computational methods, in particular virtual screening of large databases of small molecules are commonly used today to find novel leads in the process of drug development. Both a fast and accurate scoring function is required to evaluate thousands of molecules and produce a ranking list where correct solutions would be placed in top positions. Here we report the development of a new knowledge-based statistical potential to predict protein-ligand interactions. By difference with previous approaches based on distance-dependent pair-potentials [1], our scoring function has been constructed based on 3D distributions of atomic contacts found around each type of protein amino acid [2] in a large number of protein-ligand complexes extracted from the Protein Data Bank (PDB). The obtained potentials were extensively tested in docking simulations performed for hundreds of protein-ligand complexes using the program DOCK, modified to include the new scoring function.

## 2. Preparation of a Large Ligand-Protein Complex Database for Data Mining and Docking Simulations

The whole PDB was processed using our own program to identify and extract small ligands complexed to proteins and make a detailed characterization of the protein-ligand interactions in terms of inter-atomic contacts. Entries with a resolution above 2.5 Å were excluded. Fragments having more than 10 heavy atoms and peptides having less than 8 amino acids were selected. Five atom types were defined on top of the chemical elements: *hydrophobe*, *aromatic*, *donor*, *acceptor*, and *hydroxyl*, being assigned using the well-known program Babel-1.6, which was modified to translate its internal atom types into the newly defined types.

The obtained protein-ligand complexes were then filtered in several steps to create a “non-redundant” database, understanding by this that the database should not contain entries for which both the structures of the ligand and the protein binding site are similar. As result, about 3 700 *complexes* were collected in the database.

## 3. Collecting Ligand Atom–Amino Acid Contacts on Model Amino Acid Structures



**Fig. 1.** 3D distribution of protein-ligand contacts collected around reference amino acids for Arginine and Tyrosine. A ring of oxygens can be seen around the guanidine group of arginine. For tyrosine, a stacking of carbon atoms on the aromatic ring and clouds of oxygen on top of the hydroxyl group are observed.

Ligand atoms contacting protein residues in ~ 3000 complexes were collected on reference amino acid structures, as described in [2], using internal-coordinate reference systems created for every atom in each of the twenty protein amino acids, in order to deal with the conformational variability of the amino acid side chains and backbone. The resulting interaction patterns are illustrated in Fig. 1 for arginine and tyrosine.

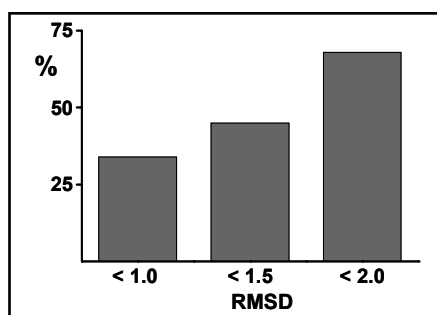
#### 4. Creating a Scoring Function from the Collected Contact Patterns

A set of scoring functions (for each atom type and each of the twenty amino acids) was created from the collected interaction patterns. For each reference amino acid the contacting clouds of ligand atoms were mapped onto its molecular surface, and the obtained contact densities were then converted into scoring terms defined across the amino acid surface. The position of grid points on the molecular surface is defined using the same internal-coordinate reference systems employed to collect the contacts.

For a given protein, a set of scoring grids is created in the binding site region, to be used within the program DOCK. Each grid point at the protein binding site receives contributions from its neighbouring amino acids by transposing the energy scores defined in the corresponding reference surface grids. A smoothing function is used to fill up a layer of grid points with a thickness of ~ 2 Å on the protein surface.

#### 5. Docking Simulations to Calibrate and Test the New 3D Statistical Potential

To test the performance of the 3D statistical potentials and adjust a few parameters, a series of automated DOCKing simulations were performed for about 600 of protein-ligand complexes from the experimental set employed for data mining. Afterwards, the potentials were tested on an independent set of about 300 complexes. We modified the program DOCK to include the set of 3D statistical potentials for scoring. All the processing of the input and output from the docking runs was done automatically. The ATPTS (“attached points”) representation of the binding site developed previously by us [2] was employed for ligand orientation.



**Fig. 2.** Percentage of protein-ligand complexes showing docking solutions with RMSD from crystal structure below 1.0, 1.5 and 2.0 Å, respectively, within the top ten solutions

#### 6. Conclusions

We compiled a large database of small ligand-protein complexes derived from the PDB, that includes information on atomic contacts and ligand atom types. This database is a valuable resource for data mining and testing of docking algorithms. From this database we extracted geometric and chemical patterns of interaction between protein amino acids and atoms from small ligand molecules, that were then converted into a 3D statistical scoring function, that takes into account the spacial position of the ligand atoms with respect to protein amino acids. This new scoring function, as incorporated into the program DOCK, was extensively tested in docking simulations performed for hundreds of protein-ligand complexes, showing very good results.

#### References

- [1] H Gohlke and G Klebe. Statistical potentials and scoring functions applied to protein–ligand binding. *Curr Opin Struct Biol* 11:231–235, 2001.
- [2] E. Moreno and K. León. Geometric and chemical patterns of interaction in protein-ligand complexes and their application in docking. *Proteins*, 47:1-13, 2002.