Statistical Potentials on Protein Structure

Jorge Roel, Cristina Leal

Universitat Pompeu Fabra

Python and Structural Biology courses, year 2015

Introduction

During the last two decades, knowledge-based potentials based on pairwise distances, also called potentials of mean force (PMF), have become a central axis for the prediction and the design of protein structure as well as for the simulation of protein folding. These PMFs are thought to be approximations of free energy functions inspired in the statistical physics of fluids developed by Sippl in 1990.

Despite this fact, PMFs are widely used with considerable success; not only for protein structure prediction, but also for quality assessment and identification of errors, fold recognition and threading, molecular dynamics, protein-ligand interactions, protein design and engineering, and prediction of binding affinity (Thomas Hamelryck et al., 2010).

Statistical potentials are energy functions derived from an analysis of a known protein structure and can be classified by the following characteristics: (1) protein representation (e.g., centroids of amino acid residues, Ca=Cb atoms, and all atoms), (2) the restrained spatial feature (e.g., solvent accessibility, contact, distance, torsion angles), and (3) the reference state (Min-Yi Shen and Andrej Sali, 2006).

Most of the statistical potentials employ the Boltzmann law to convert the observed frequencies of interactions into potentials. These potentials are obtained as the ratio of observed and expected frequencies, where the expected frequencies are derived from a hypothetical reference state when no interactions take place (Dmitry Rykunov and Andras Fiser, 2010).

However, the most essential question in the derivation of a statistical potential is how to formulate and interpret the scoring function derived from a sample of native structures. The principle of the

corresponding assumptions is that the distribution of distances between different residue types obey the Boltzmann statistical mechanics.

The aim of this project is the development of a software capable to determine such statistical potentials based on these potentials of mean force, as well as the analysis of protein residue-level structural properties..

Reference state frequencies construction and protein domains mining

Most of the statistical potentials calculated nowadays are pairwise potentials. These pairwise potentials are frequently used in combination with other types of potentials for a better protein structure assessment. Despite this, and for sake simplicity, our current work has focused just on pairwise potentials due to the high computational cost to apply the other ones.

Protein structures in the Protein Data Bank (PBD) provide a wealth of data about the interactions that determine the native state of proteins. Using the probability theory, it is possible to derive an atomic distance-dependent statistical potential from a sample of native structures.

The sampling data was retrieved from the SCOPe database (Structural Classification of Proteins) and consists of 1,110 protein-domain structures. By excluding protein-domains above a 40% homology threshold, redundancies were removed, and structures are less likely to be repeated and hence over sampled.

On the other hand, the use of protein-domains instead of whole/entire protein structures, led us to minimize even more these redundancies, as known domain structures are just one time parsed.
The reference state will reflect the probability of generating a set of pairwise distances using local structure information alone.

For the definition of the reference state we have calculated the expected residue-pair frequencies without neither taking into consideration interactions nor spatial distributions. In order to calculate the interatomic distances we have used the Cb, as it is the one that gives us information about the direction and orientation of the side-chains, instead of using the Ca.

Through the implementation of the function "*get_normalized_pairs*" (see module *get_normalized.py*), we generated a repository of residue-pair frequencies in whose Cb were below a given radius distance expressed in angstroms (Å). The frequency of a pair of residues at a distance 'r' is different if the residues are close or distant along the sequence. Concretely, we have generated five dictionaries with a given radius distance (25, 20, 15, 10, and 5Å, respectively) for future uses.

Methods

In this section we discuss the mathematical approach we have used for the calculation of statistical potentials based on the potentials of residue pairwise distances.

Many textbooks present these PMFs as a simple consequence of the Boltzmann distribution applied to pairwise distances between residues of a protein. The previously repository of frequencies represents P(r), where r is a vector of pairwise distances. This distribution, applied to a specific pair of amino acids is given by:

$$P(r) = \frac{1}{Z} e^{-\frac{F(r)}{kT}} \tag{1}$$

where r is the distance, k is the Boltzmann constant, T is the temperature, F(r) is the free energy associated to the pairwise system, and Z is the partition function.

By a simple rearrangement, and inversion of Boltzmann formula we express the energy F(r) as a function of P(r):

$$F(r) = -kT \ln P(r) - kT \ln Z \tag{2}$$

To build up a PMF we introduced the reference state with its corresponding distribution (Q(r)) and partition function (Z), so for calculating the difference of energies:

$$\Delta F(r) = -kT \ln \frac{P(r)}{Q_R(r)} - kT \ln \frac{Z}{Z_R} \tag{3}$$

As the reference state is a hypothetical system in where there are no specific interactions between the amino acids, the second term Z and Z_R can be ignored as they are constants.

Then, P(r) is the conditional probability of finding the Cb atoms of two residues at a given distance *r* from each other, giving rise to the free energy difference. The total free energy difference of a protein is then the sum of all the pairwise free energies:

$$\Delta F_{\mathrm{TOT}} = \sum_{i<j} \Delta F(r_{ij}|a_i,a_j) \quad = -kT \sum_{i<j} \ln \frac{P(r_{ij}|a_i,a_j)}{Q_R(r_{ij}|a_i,a_j)} \tag{4}$$

in where the sum of all amino acid pairs is a_i,a_j and r_{ij} is the corresponding distance threshold.

Therefore, we simplify the above expression in:

$$\Delta F_{TOT} = -kT \sum \frac{N_{obs}}{N_{exp}} \tag{5}$$

relating the frequency of a given residue-pair observed at a single protein with the same frequency at the repository previously made (see previous section).

Clearly, low free energy differences will indicate that it is more likely in the query protein than in the reference state.

Development and implementation of 'statools' software

As above mentioned, we perform the calculus based on Cb distances, so the first step is to capture those atoms by the function "*parse_atoms_infile*" (See module *functions.py*); which returns a list with all the Cb atoms present in a pdb style file (files with extensions .pbd or .ent).

Even though we are not interested in the whole list of Cb atoms, we still need to get those residue pairs, whose Cb atoms are below a given radio distance. We achieve this purpose by the implementation of the function "*get_neighbors*", that applies a nearest-neighbour search (NN) algorithm at a given radius by the use of a C+ implemented Kd-Tree, and returns a list of the atom pairs we are interested in (see module *functions.py*).

What we need now is residue-pair frequencies (instead of atoms) with their corresponding positions in the protein (function "*get_infile_pairs*") in order to compare them to the repositories previously created (see module normalized_pairs.py), for the following calculus.

Now by applying equation (5), (see previous section), in where N_{obs} is the frequency of a given pair observed in the target protein, N_{exp} is the frequency of the same pair at the repository, k is the Boltzmann constant (0.0019872041 kcal/mol), and T is the temperature (297.15 K), we obtain the statistical potential of that residue-pair in each specific position of the protein structure.

By definition, the statistical potential for a given residue (at a given position), as aforementioned, is the sum of all the pair potentials taking place at such position. So, by a simple addition of these potentials, we obtain the total statistical potential for a given position (corresponding to a specific residue) in a protein.

Finally, we plot an energetic protein profile, which contains each individual residue statistical potential. Y axis represented in terms of energy (kcal/mol), and X axis represented in terms of position-residue (see module *plot.py*).

Tutorial examples

This section is thought to be a kind of handbook about the utilities of the 'statools' software.

Installation:

To build and install Stats, download and unzip the source code, go to this directory at the command line, and type:

```
python setup.py build
sudo python setup.py install
```

Please, see README.txt for further information.

Usage:

stats [-h] [-i INFILE] [-o [OUTFILE]] [-r [RADIUS]] [-v]

-i [INFILE], --input [INFILE]

Input file. Input file needs to have a PDB Format containing protein atoms and its coordinates.

-o [OUTFILE], --output [OUTFILE]

Outfiles the Position-Residue and its associated statistical potential. If not defined, results are printed to standard output.

-r [RADIUS], --radius [RADIUS]

Defines the desired radius of neighbor-search algorithm. If not set, the default radius given is 10Å. Radius has to be set to either: 5, 10, 15, 20, or 25Å, otherwise default value will be used.

-v, --visual

If this argument is defined, results are plotted and saved as '*filename.png*' in the working directory. If not used, plot is not showed.

By typing at command line level after installation:

stats -i 1EPG.pdb -o 1EPG.pdb.out -r 10 -v ;

stats program generates two different files. On one hand, a plane text file corresponding to the statistical potential of each residue at the given radius, "*1EPG.pdb.out*", and on the other hand an image containing the plot with the protein energetic profile, "*1EPG.pdb.png*". Both outputs and input provided at STATS directory.

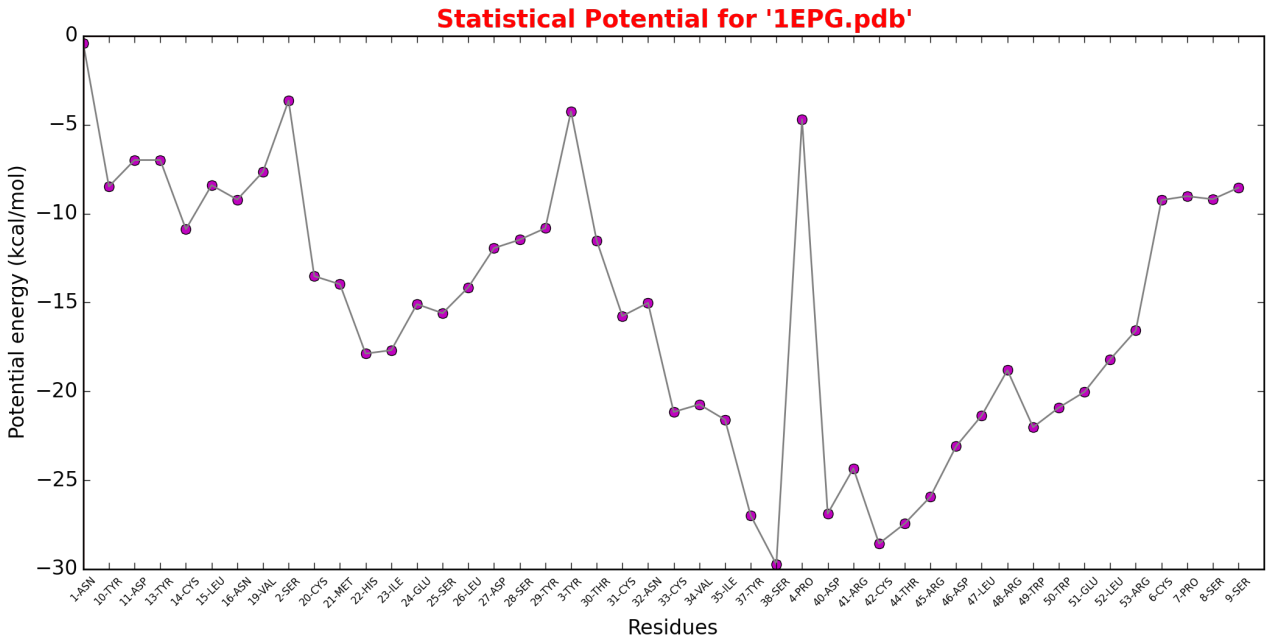

Figure 1. Energetic profile of protein 1EPG structure. Y axis potential energy (kcal/mol), X axis residues.

Discussion

The statistical distributions of residue-level distances in known protein structures provide a valuable source of information for estimating these residue level structural properties of proteins, which are not otherwise accessible experimentally. However, this statistical measure relies upon the quality as well as on the quantity of sampled known structures.

We have downloaded around one thousand structures from the PDB Data Bank, which should be enough to obtain reliable statistical estimates of the distribution of pair-residues according to a given radius, and in such way building a ‘reference state’.

In our software ‘stats’ we apply the above described methodology and aims to estimate the statistical potential for any protein structure in PDB format (‘.ent’ or ‘.pdb’), but further refinements should be implemented in order to make them computationally meaningful (such as solvation energy).

The use of the tool of this study is a energy-position plot for correlations between pairs of neighbouring residues. An example has been proposed above in the present report (see Tutorial examples).

To sum up, with the increase in both number and quality of known protein structures, many structural properties can be derived from sets of protein structures by statistical analysis and data mining, and these can even be used as a supplement to the experimental data for structure determinations.

References

Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, et al. (2010) *Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized.* PLoS ONE 5(11): e13714. doi:10.1371/journal.pone.0013714

Hui Lo, Long Lu, and Jeffrey Skolnick (2003) *Development of Unified Statistical Potentials Describing Protein-Protein Interactions*. Biophysical Journal vol.84

Wu, Di, (2006) *Distance-based protein structure modeling*. Retrospective Theses and Dissertations. Paper 3036.

Li Y, Liu H, Rata I, Jakobsson E (2013) *Building a Knowledge-Based Statistical Potential by Caputring High-Order Inter-residue Interactions and its Applications in Protein Secondary Structure Assessment*. J. Chem. Inf. Model. 2013, 53, 500−508

Rykunov D, Fiser A (2011) *New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics, 11:128.* http://www.biomedcentral.com/1471-2105/11/128