PRInS 1.0

Pavlos Pavlidis

Introduction

The goal of *PRInS* is to detect residues in a protein with a significant functional role. *PRInS* implements a simple statistical idea, that can be summarized as follows: Residual-interactions that are statistically rare are preserved during evolution only if there is an important fitness-related reason able to counteract the fact that the interaction is unfavorable. In other words, some residual-interactions are observed rarely because the interactions are not favored. If a certain residue has a sufficient number of those interactions, it means that there should be some reason that this residue has been preserved during evolution. For example, it might be that the specific aminoacid has some important functional role and thus, it cannot be substituted by another residue.

PRInS is developed in C programming language and is available as an open-source software under the GPLv3.0 license.

For queries, bugs reports etc please contact:

Pavlos Pavlidis
email: pavlidisp@gmail.com

Command Line

Command Line Options

```
prins -h | --help: Prints out this help

-iMatrix <FileName>: imports a matrix that describes
statistical interactions

-oMatrix <FileName>: outputs the statistics for the
interactions between residuals in a file, in a matrix form

-pdb <FileName>: reads in the pdb file

-o <FileName>: outputs result to a file

-list <FileName>: processes a list of files. Their filenames
are written in a list, one name in each line

-threshold <float>: distance in Angstrom to consider two
neighboring Ca atoms as interacting (default: 6.5)

-membrane: analyze atoms only within the membrane
boundaries

-rfile <file name>: print out statistics from permutations of
the aminoacids in protein structures

-nrand <int>: perform <int> randomizations (permutations) of
the aminoacid chain (default: 0, i.e. don't do randomizations)
```

Details on the command line options

-h

`-h` prints out the help for the command line, i.e. short information for each flag in the command line

-iMatrix

Using the flag `-iMatrix` it is possible to import an already calculated matrix describing the statistics for the aminoacid interactions. **For the moment**, please use only matrices describing the interactions for environments I, II, and III, as well as the interactions within each environment. Such a matrix corresponds to the Scoring Matrices for the *Residue-wise contact-based environment* in the Jha et al. [2011] study.

-oMatrix

Using the flag `-oMatrix` you can write in a table the matrix describing the statistics of the interactions between the residues. Again, **for the moment** this only corresponds to the *Residue-wise contact-based environment* in the Jha et al. [2011] study.

-pdb

Load a structure, i.e. a pdb file. If you use this option, most probably you would like to use the `-iMatrix` flag as well, to specify the input matrix.

-list

In contrast to the `-pdb` flag, you can specify a **set** of pdb files within another file that contains just the filenames of the pdb files. In other words, if you would like to analyze the structures specified in the files: 1EZS.pdb, 2EZR.pdb and 2ZZR.pdb, then you should create another file, e.g. all_files.txt, and type as contents the filenames of the files you want to analyze, i.e.:

```
1EZS.pdb
2EZR.pdb
2ZZR.pdb
```

-o

Using this flag you should specify an output file, where the results will be written to.

-threshold

With `-threshold` option, you set the distance that determines whether two residues are considered neighbors (or interacting). Here, we measure by default the distance of the C_a atoms of the residues. The default is 6.5 Angstrom. This means that if the distance between the C_a of residues **A** and **B** is smaller than 6.5 Angstrom, then the residues **A** and **B** are considered to be neighbors (or interacting).

-membrane

Often, we have information about the location of the cellular membrane. For example, the database at `http://opm.phar.umich.edu/` provides information for the location of the membrane assuming alpha-helix proteins.

-nrand

With the `-nrand` option you can specify the number of permutations for the randomization tests. In fact, what we do here, is that we calculate a score for the whole pdb file. This score is given by the formula:

$$z = \frac{\mu_{rand} - E_n}{\sigma_{rand}} \tag{1}$$

where μ_{rand} is the mean score of a random permuation of all residues in the pdb file, E_n is the score of the real protein, and σ_{rand} is the standard deviation of the scores of the permutated sequences.

-rfile

Using the `-rfile` flag, you can specify a file where to print out the statistics for the permutated sequences. In this file, the score of each permutated sequence will be printed out. Then, you can import this file in R and do some fancy statistics (i.e. histograms of scores etc).

Examples

PRInS package contains a folder named 'data'. There is an executable file, run.sh, which contains instructions (commands) that download and analyze the 83 protein structures used in the Jha et al. [2011] study. To execute the script just do:

- `cd data`
- `./run.sh`

You will need internet connection to run the script properly. The reason is that the protein structures are downloaded by the script.

Bibliography

Anupam Nath Jha, Saraswathi Vishveshwara, and Jayanth R Banavar. Amino acid interaction preferences in helical membrane proteins. *Protein engineering, design & selection : PEDS*, 24(8):579–88, August 2011. ISSN 1741-0134. doi: 10.1093/protein/gzr022. URL `http://www.ncbi.nlm.nih.gov/pubmed/21666247`.