MyPMFs – User manual

Version 1.0; 2018-04-09

G. Postic, T. Hamelryck, J. Chomilier, D. Stratmann

1. Installation

Simply type `make` in the terminal.

This will create the two executable binaries, named `scoring` and `training`, which compose MyPMFs.

2. Options

Below are listed the options of the `scoring` and `training` executable binaries.

Note: When run with forbidden values or incompatible options, these programs will return an error message.

2.1 `training` options

`-l` [character string]

The input file (name and path) of the list of PDB codes used for the training of the statistical potential.

The list must contain one PDB code by line. The fifth character is facultative and represents the chain name; if not provided, every chain found in the coordinate file will be processed.

Example:

```
1A3C
1A4IA
1A62
1A8D
1AH7A
1AHO
1AMT
```

`-d` [character string]

The directory containing the PDB files listed in the -l input.

Three formats are accepted for the PDB file names:

```
xxxx.pdb
XXXX.pdb
pdbxxxx.ent
```

where `x` and `X` correspond to lower- and uppercase characters, respectively.

`-L` [character string]

Alternative input to `-l` (and `-d`); for training on non-native protein structures.

The input file (name and path) of the list of coordinate files used for the training of the statistical potential.

The list must contain one filename (including path) by line. The chain name may be optionally added, separated by a space.

Example:

```
path_example/dataset/1A3C.pdb
path_example/dataset/1A4I.pdb B
path_example/dataset/1A62.pdb
path_example/dataset/1A8D.pdb
path_example/dataset/1AH7.pdb
path_example/dataset/1AHO.pdb
path_example/dataset/1AMT.pdb C
path_example/dataset/1ATG.pdb
```

`-o` [character string]

The output directory where every file generated by the `training` executable will be stored.

It can be either an existing or non-existing directory (the latter would then be created).

`-m` [floating-point number]

The upper distance threshold (Å), above which the interatomic distances, observed in the input coordinate files, are not taken into account. The default value is 15.0.

`-n` [floating-point number]

The lower distance threshold (Å), below which the interatomic distances, observed in the input coordinate files, are not taken into account. The default value is 0.0.

`-w` [floating-point number]

Width (Å) of each bin in the distance distributions. The default value is 0.1.

`-x`

Will only take into account the inter-chain interactions, *i.e.* the pairwise interatomic distances between two atoms belonging to different subunits of a multimeric protein structure.

This option is aimed at generating statistical potentials for the scoring of docking poses.

`-y`

Will only take into account the intra-chain interactions (see `-x` above).

By default, the `training` program takes into account both intra- and inter-chain interactions.

`-i` [integer number]

Every pair of atoms taken into account for calculating the distance distributions will be separated, on the protein chain, by a minimum number of residues defined by `-i`. The default value is 3.

`-j` [integer number]

Maximum number of residues separating the interacting pair (see `-i` above). By default, this number of residues is not limited.

`-t` [floating-point number]

For two interacting atoms, separated by a distance d, this option defines the maximum value of the pseudo-energy. The default value is 10.0 (arbitrary units).

`-r` [character string]

The `training` executable allows generating statistical potentials using either every atom type of the protein structures from the training set, or simpler representations (*i.e.* using only some atom types). The program can also handle coarse-grained (CG) models. The possible values are: `CA` (Cα-only), `CB` (Cβ-only), `BB` (backbone beads), `backbone`, `sidechains`, `SC1` (side chain beads 1), `allatom`, or `allatomCG` (BB, SC1, SC2, SC3, and SC4), `sidechainsCG` (SC1, SC2, SC3, and SC4). The default value is `CA`.

`-g`

For every pair of atom, `training` will write a `.dat` file containing all the interatomic distances. This option is automatically activated when using kernel density estimations.

`-p`

This option is used for plotting the distance-dependent pseudo-energy profiles for every pair of atoms. This produces vector images (SVG format). When combined with kernel density estimations, the histograms and densities of all atom pairs will also be plotted.

`-f` [character string]

The path to the directory where the required R scripts are located. The default value is `./src/`.

`-A`

When this option is activated, the distances are not stored into the RAM memory (the distance distributions are calculated 'on-the-fly'), which allows processing large datasets of protein structures. However, this option is not compatible with the use of kernel density estimations.

The following three options aim at training the reference state separately from the observed frequencies (e.g. training using different parameters).

`-W`

This option writes the frequencies of the reference distribution into a file which will be named `frequencies.ref`.

`-R` [character string]

Path to the `frequencies.ref` file for generating a statistical potential using a pre-computed reference state.

Example for training the reference state using all-atom representations, and then training the observed frequencies on Cα only (`training` is therefore used twice):

```
./training -l example/list1.txt -d example/dataset/ -o myReference
-r allatom -W
```

```
./training -l example/list1.txt -d example/dataset/ -R
myReference/frequencies.ref -r CA -o myPotentials
```

`-X`

Same as `-W`, but for large input datasets; not compatible with the use of kernel density estimations.

The following three options also aim at training the reference state separately from the observed frequencies, but when using a set of random decoys.

`-Y`

The frequencies of the distributions that will be used for the reference state are written into one `.frq` file for each atomic pair.

`-Z` [character string]

Directory containing the `.frq` files for generating a statistical potential using a pre-computed reference state.

Example for training the reference state using a list of random decoys, and then training the observed frequencies on native structures:

```
./training -l example/decoys_list.txt -d example/random_decoys/ -o
myReference -Y
```

```
./training -l example/list1.txt -d example/dataset/ -Z myReference/
-o myPotentials
```

`-U`

Same as `-Y`, but for large input datasets; not compatible with the use of kernel density estimations.

The `training` executable of MyPMFs has also six options related to the `density()` function from the R standard library (kernel density estimations).

For more details, please refer to the R documentation: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/density.html

2.2 `scoring` options

`-i` [character string]

A single input coordinate file (PDB format).

`-a` [character string]

Chain name (when using `-i` option); e.g. AbC will process chains A, b and C (default: all chains)

`-l` [character string]

Alternative input to `-i`.

The input file (name and path) of the list of coordinate files to be scored with the statistical potential.

The list must contain one filename (including path) by line. The chain name may be optionally added, separated by a space.

Example:

```
path_example/dataset/1A3C.pdb
path_example/dataset/1A4I.pdb B
path_example/dataset/1A62.pdb
path_example/dataset/1A8D.pdb
path_example/dataset/1AH7.pdb
path_example/dataset/1AHO.pdb
path_example/dataset/1AMT.pdb C
path_example/dataset/1ATG.pdb
```

`-d` [character string]

Directory containing the pseudo-energy files `.nrg` generated by the `training` program.

`-o` [character string]

The output directory where every file generated by the `scoring` executable will be stored.

It can be either an existing or non-existing directory (the latter would then be created).

`-m` [floating-point number]

The upper distance threshold (Å), above which the interatomic distances, observed in the input coordinate files, are not taken into account. The default value is defined by the `parameters.log` file contained in the `-d` directory.

`-n` [floating-point number]

The lower distance threshold (Å), below which the interatomic distances, observed in the input coordinate files, are not taken into account. The default value is defined by the `parameters.log` file contained in the `-d` directory.

`-x`

Will only take into account the inter-chain interactions, *i.e.* the pairwise interatomic distances between two atoms belonging to different subunits of a multimeric protein structure.

`-y`

Will only take into account the intra-chain interactions (see `-x` above).

By default, the `scoring` program takes into account both intra- and inter-chain interactions.

`-j` [integer number]

Maximum number of residues separating the interacting pair (see `-k` below).

`-k` [integer number]

Every pair of atoms taken into account for calculating the distance distributions will be separated, on the protein chain, by a minimum number of residues defined by `-i`. The default value is defined by the `parameters.log` file contained in the `-d` directory.

`-q` [character string]

Comma-separated list of residues (number+chain) that will be processed (default: all residues).

Example:

```
./scoring -i example/dataset/1BKR.pdb -d myPotentials/ -q
10A,11A,12A,13A,14A,15A,16A,17A,18A,19A,20A
```

`-r` [character string]

Representation: `CA` (Cα), `CB` (Cβ), `BB` (backbone beads), `backbone`, `sidechains`, `sidechainsCG`, `SC1` (side chain beads), `allatom`, or `allatomCG`. The default value is defined by the `parameters.log` file contained in the `-d` directory.

`-c`

A cubic spline interpolation of the discrete potentials is used for calculating the pseudo-energy of the query structure (by default: linear interpolation).

`-p`

Use this option for plotting the distance-dependent pseudo-energy profiles for every pair of atoms. This produces vector images (SVG format).

`-w`

Write files `energy_[window size].tsv` (energy /position) and `data.tsv` (energy and distance /atom pair). See `-b` below.

`-b` [integer number]

Window size for calculating the energy per position (default=1)

`-f` [character string]

The path to the directory where the required R scripts are located. The default value is `./src/`.

The following four options aim at calculating a Z-score of the pseudo-energy. Please refer to the supplementary material of the MyPMFs article (Postic *et al.*, 2018, *Biochimie*) for more details.

`-z`

Z-score computation (only for Cα and backbone beads representations)

`-M`

Mute the counter of random sequence decoys (important when redirecting the output)

`-s` [integer number]

Number of random sequence decoys for the Z-score computation (default=1000)

`-t` [floating-point number]

For each random sequence decoy: maximum sequence identity with the query structure (default=0.5)