Laboratory of Bioinformatics and Computational Intelligence

labic.utfpr.edu.br/

Protein structure prediction analysis using the biological model

Step-by-step:

The purpose of this tutorial is to describe in a simplified way how to generate your protein analysis files . The steps required to accomplish this goal are presented below.

step1) Generate your protein AB sequence and the biological model using alpha carbons data from PDBx/mmCif file
step2) Calculate statistics informations about the distance between each residue
step3) Metrics calculation: F1-score, Recall, Precision, and Matches, comparing the real model with the predicted one
step4) Generate a .pdb file with all the coordinates to 3d model, formatted specifically for the RasMol program

###
step1) Generate your protein AB sequence and the biological model using alpha carbons data from PDBx/mmCif file
###

1.1) Download the intended affinity hydrophilic scale and put into the “data_input” directory.

1.2) The Alberts scale is set as default in the executable so if necessary to change the scale just open the “exec.sh” and the “exec_pdb3.sh” files and change the “table” variable from “alberts” to the intended scale.

1.3) Execute the “pdb3.py” file in the “src” directory as python3, example:
“python3 src/pdb3.py”

1.4) Put the intended protein in the terminal, in lowercase, and the name of the intended scale.

1.5) This program will generate the following files, where “protein” is the name of the protein:

output:

1. *seq_"protein".txt* **-** List of the amino acids sequence

2. *"protein”_A.pdb* - List of all atoms in the protein, with their coordinates and the amino acid that they belong

3.*proteins.txt* - List of all the proteins intended to use in the other programs

4.*"protein".ciff* - File with all the data about the protein is extracted from

5.*"protein"_CA.txt* - List of all alpha carbons in the protein, with their coordinates and the amino acid that they belong

6.*seq_10_"protein".txt* - Sequence of the hydrophilic affinity formatted for easier manipulation

7.*seq_ab_"protein".txt* - Sequence of the hydrophilic affinity formatted for easier understanding

8.*posicao_com_numeracao_"protein".txt* - List of alpha carbon hydrophobics atoms, with their coordinates and position in the protein

9.*"protein"_pos_CA_h.txt* - List of alpha carbon hydrophobics atoms, with their coordinates and position in the protein, formatted for easier manipulation

##

step2) Calculate statistics informations about the distance between each residue

##

2.1) Execute the “pdb3.py” file, or the “exec_pdb3.py” executable, to generate the necessary input:

input

1.proteins.txt
2."protein"_pos_CA_h.txt

2.2) Execute the “calc_statistics.py” file in the “src” directory as python3, example: “python3 src/calc_statistics.py”

2.3) This program will generate the following files, where “protein” is the name of the protein:

output

1.*dist_euc_all_prot.txt* - Sequence of all Euclidean distances, between hydrophobics residues, of all proteins obtained in the “pdb3.py” program.

2.*"protein"_euc_dist.txt* - Matrix of the Euclidean distances between the hydrophobic residues of a single protein, where [1][0] = [0][1].

3.*all_euc_dist_asc.txt* - Sequence of all Euclidean distances, between hydrophobic residues, of all the proteins obtained in the “pdb3.py” program, formatted for easier visualization.

4.*statistcs.txt* - Statistical information, such as mean, standard deviation, maximum, second to maximum and minimum of the Euclidean distance. The calculations take all the proteins inserted into consideration.

###

step3) Metrics calculation: F1-score, Recall, Precision, and Matches, comparing the real model with the predicted one

###

3.1) Execute the “pdb3.py” file, or the “exec_pdb3.py” executable, and “calc_statistics.py”, or the “exec_calc_statistics.py” executable, to generate the necessary input:

input

1.pred.csv - the Predicted model used to compare the biological structure
2.proteins.txt
3."protein"_euc_dist.txt
4.output_protein_data_"protein".txt
5."protein"_CA_euc_map.txt

3.2) Execute the “ver_pos_score.py” file in the “src” directory as python3, example: “python3 src/ver_pos_score.py”

3.3) This program will generate the following files, where “protein” is the name of the protein:

output

1.*protein_matchs.csv* - A table with the number of true positives given a threshold set.

2.*protein_f1scores.csv* - A table with F1-scores measure given each threshold.

3.*protein_recall.csv* - A table with the Recall calculations given each threshold.

4.*protein_precision.csv* - A table with the Precision calculation given each threshold.

5.*Dists_x_Score.csv* - A table of the score, the difference of the expected contacts number and the contacts number of the predicted model in each threshold.

##
step3) Generate a .pdb file with all the coordinates to 3d model, formatted specifically for the RasMol program
##

4.1) Execute the “pdb3.py” file, or the “exec_pdb3.py” executable, to generate the necessary input:

input

1.proteins.txt

2.output_protein_data_"protein".txt - Predicted model used to comparison with the biological protein model

4.2) Execute the “create_3d_model.py” file in the “src” directory as python3, example: “python3 src/create_3d_model.py”

4.3) This program will generate the following files, where “protein” is the name of the protein:

output

1.*coordenadas_rasmol_"protein".pdb* - A file with the coordinates for a AB off-lattice 3D model, formatted for the RasMol program.