Algorithms in Structural Bioinformatics
Motifs Search Practical

Homework

As you may know, functional motifs can be discovered in DNA or protein sequences. The objective of this practical is to use the 3D information of protein structures and combine it with sequence information to see whether this can help the discovery of functional motifs. In this practical you will manipulate PDB files with *R* software.

Exercice - 1 Sequence and structute similarity

Retrieve the archive `http://www.lgm.upmc.fr/laine/STRUCT/TD/TD-03_data.zip`. It contains the coordinate PDB files of two proteins and their amino acid sequences in fasta format.

1- Use the program *Stretcher* (for example from the Mobyle portal `http://mobyle.pasteur.fr/`) to perform global alignment of the two protein sequences. What percentage of identity do they share?

2- Use the program *SSM* (`http://www.ebi.ac.uk/msd-srv/ssm/`) to perform structural alignment of the two protein tertiary structures. What are the values of the Z-score and the RMSD? Download the aligned structures and visualise them with *Pymol*.

3- In your opinion, are these two proteins homologous?

Exercice - 2 3D information encoding and 3D motif search

1- Write an R function *createSeq3D* that generates a "3D-sequence" representing a protein, given its amino acid sequence and tertiary structure. The function will:

a- read a PDB file and get the sequence of amino acid residues
b- compute the inter-residue distance matrix where each element (i,j) is the minimal distance between residues i and j over all their atoms
c- compute a contact matrix where each element (i,j) is the minimal distance between residues i and j over all their atoms **if the residues are linked in 3D, zero otherwise**; A 3D link will be created between two residues if they are separated by at least 7 residues in the primary structure and by at most 3.6 Å in the tertiary structure
d- transform the sequence of the protein into a "3D-sequence" that encodes 3D-links. At each position i in the sequence you will consider the residues that are linked in 3D (3D-neighbors) with residues i and you will add them in the order of their increasing distances to residue i. If two (or more) 3D-neighbors of residue i are separated by less than 8 residues in the primary structure of the protein, then only the one that is at the smallest distance to residue i will be added next to it.
e- return the "3D-sequence" representing the protein along with the positions of the residues in the original 1D sequence.

You may use the *bio3d* package: `http://thegrantlab.org/bio3d`.

2- Write an R function *getMotifs* that detects motifs of a given length l shared by two "3D-sequences" given as input. The function will retain only motifs that are not purely based on sequence information, *ie* motifs containing residues that directly follow each other in the original 1D sequence will be discarded. The function will return the starting positions of the motifs in the two "3D-sequences". You may use the *words.pos* function of the *seqinr* package.

Exercice - 3 Functional motif characterization

1- Apply the R functions you developed to the two proteins 1ACB_E and 1DNQ_A.

a- Generate the "3D-sequences" of the two proteins
b- Extract motifs of length 4 common to the two "3D-sequences"

2- One of the extracted motifs contains a triad of residues crucial for the catalytic activity of the two proteins. Can you find out which triad is that?

a- Go look at the corresponding PDBsum pages (`http://www.ebi.ac.uk/pdbsum`) to determine the name of the two proteins, and the organisms to which they belong.
b- Perform a bibliographic search to try and find more information of these two proteins, in order to identify the triad of residues.

3- Could you have identified this motif based on the sequence and structure alignments? Can you think of another application of the method proposed here? How would you improve the construction of the "3D-sequences"?