1 Protein part

Your task is to examine the local structure of protein segments with an identical amino acid composition.

The same amino acid sequence can show up in different proteins. For example, the sequence ACAFT might be found in protein A and in protein B. However, as we saw, that same sequence might adopt different local structures in protein A and in protein B. For example, in protein A, the segment might be a helix, while in protein B the segment might be a sheet. You will examine this kind of variability using the RMSD measure.

Your detailed task is to:

1. Find all cases where the same sequence of 5 amino acids occurs in two proteins, using a database of proteins (see below). Try to find an efficient way to do this, for example using a Python dictionary. Make sure to check for chain breaks: consecutive CA atoms should be within 4 Å from each other.
2. For all matching pairs, calculate the RMSD after optimal superposition using the SVD-based algorithm discussed in the course. Implement this algorithm yourself - do not use the one in Bio.PDB or elsewhere. Use only the CA atoms for the calculation of the RMSD.
3. Make a histogram of the RMSD values. Hint: you can use Python + Matplotlib for this:
   (a) http://matplotlib.org/1.2.1/examples/pylab_examples/histogram_demo.html
4. Discuss the results. How many matching pairs did you find? Do they have the same structure? How does the RMSD plot look? Do you see any peaks? What do they correspond to?
5. Make the same plot for the same number of randomly chosen 5 residue fragment pairs from different proteins. Compare with the previous plot and discuss.

As data base, use the top100 collection of high quality protein structures:
`http://kinemage.biochem.duke.edu/php/downlode.php?filename=/downloads/datasets/top100H.tgz`

Use Bio.PDB to implement the script. Disregard any structures that cannot be parsed by the Bio.PDB parser, but ignore warnings.

2 RNA part

In this exercise you fold RNA sequences by energy minimization and compensatory base pair changes and analyze some properties in more details.

1. Consider the RNA sequence:
   AUCAGUUCUAGCAGGAGCUGUACUCAGAGACUCGGGAAAUUUUCCCGGAAUUUUACCCGGGUUUUUACGU
   and fold it in the RNA Vienna webserver:
   http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi.

   (a) Describe briefly what type of secondary structure you obtained and inspect the base pair probabilities of the structure in the dot plot.
   (b) Download the .eps file of the base pair probabilities and write a script that sorts and prints the base pairs, their corresponding position along with the probability. (Hint: the upper part of the matrix is indicated by the lines ending with "ubox". These contain the positions and the base pair probabilities). Ensure that they are sorted from highest to lowest. How many base pairs have a probability of greater than 80%?
2. Now use the same webserver and your script to generate the similar output for the sequence:
   AUCGGUUCCAGCAGGAACUGUACUCGGGGGCUCGGGAAACCCUCCCGGGGUUUUACCCGGGUUUUUACGU
   Answer the same questions as in 1a and 1b.
3. Write a Python script that computes the Hamming distance between two strings of the same length. Implement the calculation of the hamming distance yourself (do NOT use any library that does the job for you!). Use the script to calculate the hamming distance between the two sequences and the two secondary structures in dot bracket format from task 1 and 2.
4. Write a Python script to compute the base pair distance between two secondary structures in dot bracket format. Implement the calculation of the base pair distance yourself (do NOT use any library that does the job for you!). Use the script to calculate the base pair distance between the two secondary structures from task 1 and 2.

5. Discuss the essential similarities and differences between the foldings obtained on the two sequences in task 1 and 2. Consider
   (a) the two structures (hamming distance and base pair distance),
   (b) the dot plots, and
   (c) the ranked base pair probabilities.

6. The two sequences can be aligned as follows:
   AUCAGUUCUAGCAGGAGCUGUACUCAGAGACUCGGGAAAUUUUCCCGGAAUUUUACCCGGGUUUUUACGU
   AUCGGUUCCAGCAGGAACUGUACUCGGGGGCUCGGGAAACCCUCCCGGGGUUUUACCCGGGUUUUUACGU
   Submit the alignment to the RNAalifold server
   http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAalifold.cgi.
   (a) Compare the RNAalifold consensus structure (dot bracket notation) of the aligned sequences to each of the individually obtained dot bracket notations. Compute the hamming distance and the base pair distance between the consensus structure and the individual structures (dot bracket notations).
   (b) How many consistent and compensatory base pair changes exist in the consensus structure?

3 Report format

Your report consists of:

1. A PDF file for the protein part (without the code).
2. The Python code for the protein part.
3. A PDF file for the RNA part (without the code).
4. The Python code for the RNA part.

The **PDF file of the protein part** consists of:

- An introduction, with background information and an overview of the task.
- A materials and methods section, that describes the implementation and other relevant subjects (such as the data set used).
- The results of applying your methods to the dataset.
- A section with conclusions.

Note that you should provide references to the literature – NOT TO WIKIPEDIA – where needed for both the RNA and the protein part.

4 Plagiarism warning

Note that your exams will be checked for **plagiarism** by an automated method. Please do NOT work together and do NOT discuss the exam with each other. Please do NOT use verbatim quotes without proper referencing. Plagiarism might get you expelled from the university!