Exam Structural Bioinformatics 2018/19

Thomas Hamelryck & Jan Gorodkin

1 Protein part (2/3 of points)

1.1 Theory part (1/3 of points of protein part)

- *Knowledge based potentials of mean force* (PMFs) for protein structure prediction are not true PMFs in the physical sense. Discuss.
  - You can use the information from the course slides and the articles provided on Absalon, and other reliable sources you see fit (provide references).
  - Maximum 1 page.

1.2 Practical part (2/3 of points of protein part)

Your task is to examine the local structure of protein segments. Specifically, you are going to analyse the structural variability of segments of 9 amino acids in function of the central amino acid type.

- As **protein 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.

1. Extract all 9-residue polypeptide segments from the top100 database.
   (a) Exclude fragments with unusual amino acids (ie. not belonging to the 20 standard amino acid types).
   (b) Note: Make sure that the 9-residue segments are actually connected polypetides. Explain how and why you need to do this.
2. For Ala, Gly, Pro, Phe, Asp, Arg and Leu, select 500 random segment pairs that have that amino acid type as central residue (ie. as the 5th residue). The two fragments in a pair need to come from different proteins. For each pair:
   (a) Calculate the RMSD value after optimal superposition, using the SVD based method. In the RMSD method, center each fragments on the C-alpha atom of its central amino acid (the 5th residue), instead of the center of mass. Use the backbone atoms (N, C-alpha, C) of all 9 residues for the RMSD calculation.
   (b) Note: implement the optimal superposition method yourself.
3. Make a histogram of the RMSD values for each of the 7 amino acid types.
   (a) Hint: you can use Python + Matplotlib for this:
   `https://matplotlib.org/gallery/statistics/hist.html?highlight=histogram`
4. Make the same RMSD histogram for 500 superimposed 9-residue segments, randomly chosen regardless of the central amino acid type. Again, make sure the fragments come from different proteins for each pair.
5. For Proline, select the 6 fragment pairs with the lowest RMSD, visualize them in PyMOL, and put the (high quality) pictures in the report.

   (a) Hint: once you identified the fragment pair, you can superimpose them in Pymol using *pair_fit*.

      i. See `https://pymolwiki.org/index.php/Pair_fit`

6. Show the 8 RMSD histograms (Ala, Gly, Pro, Phe, Asp, Arg, Leu and the general case). Discuss the results. Which amino acids are more variable and why? Can you explain any peaks? What are the similarities and differences between the histograms? Discuss.

2 RNA Part (1/3 of points)

Micro-organisms of the genus *Bacillus* are relevant for both academic and industrial research. They are capable of performing complex regulatory mechanisms, and have tremendously different physiologies.

The tmRNA genes, which only exist in Bacteria, are involved in rescuing ribosomes that are stalled due to a skipped or a non-existing stop codon on a given transcript. The tmRNA function is carried out through a complex pseudo-knotted RNA structure.

In the file `baci_tmrna.fasta` an alignment of Bacillus tmRNAs has been extracted from a curated resource (`https://rth.dk/resources/rnp/tmRDB/rna/tmrna.html`). The first "sequence" in this file is the pairing mask describing how the different helices are made, e.g. ***helix 4*** `44444....44444` should be interpreted as `(((((....)))))`. This pairing mask is a consensus of multiple sequences expanding beyond the bacillus clade, and each sequence having upper case letters is involved in the pairing, e.g. for the entry

```
>Baci.anth._AE016879:
44444--33333------44444
gGAUcggCCUCGuuaaaacGUCa
```

only the `GAU` is base paired with `GUC` in helix 4. (Note that a part of helix 3 is ignored in this example.)

Analyze the structure and sequence as follows and write it up as described below.

- Write a script which extracts the base pairs for each individual sequence. See the explanation of a pairing mask above. Note that some sequences have gaps (symbol "-"), and the letter "n" indicating that a base is not known (and not assigned in any base pair). You may use SeqIO from biopython for parsing the input file. Hint: The first helix symbol indicates a base pair with the last symbol if the corresponding letters in the sequence are upper cases. Otherwise these two positions do not base pair in that sequence although other sequences can base pair at those positions. Use the script to find the base pairs for each sequence in the fasta file.

- Write a script that calculates the base pair distance between any pair of sequences (recall that you have to compare *pairs* of positions) and plot their distribution. Comment on your observations, such as which and what proportion of structures are close and far apart and what does that say about the relationship of the individual sequences and structures. (You may use the plotting library *matplotlib* or any other plotting method, as long as you state them in the hand-in.)

- Write a script that calculates the Hamming distance between any pair of sequences by incrementing a counter by one if the bases are different and do not increment otherwise. Ignore positions where there is (1) a gap in both sequences or (2) the letter "n" is compared to a letter different from a gap. Again, plot the distribution of the distances and do similar considerations as above. Subsequently compare to the distribution of base pair distances and comment on your observations.

3 Report format

3.1 Protein part

The written report consists of a **PDF file** (called `protein.pdf`) containing:

1. The protein theory part (max 1 page).

2. The protein practical part (max 4 pages).

   (a) An introduction, with background information and an overview of the task.

(b) A materials and methods section, that describes the theory, implementation and other relevant subjects (such as the data set used).

(c) The results of applying your methods to the dataset.

(d) A section with conclusions.

3. The Python code for the protein practical part.

   (a) Use well-structured, well-commented code.

In addition, also upload the **Python script** separately as `protein.py`.

3.2 RNA part

The written RNA report consist of a **PDF file** (called `rna.pdf`) containing:

- An introduction, including a brief background on tmRNA and an overview of what base pair distance and Hamming distances are about (max. 1 page).
- A "materials and methods" part, that describes the implementation and other relevant subjects (such as the data set used) (max. 2 pages).
- The results of applying your methods to the data set, focusing (as outlined) on the distributions to be generated (max. 2 page).
- A section discussing the suitability (weaknesses and strengths) of the approach and your conclusions (max. 1 pages).

In addition, also upload the **Python script** separately as `rna.py`.

3.3 General remarks

Please 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 WILL get you expelled from the university!