Exam Structural Bioinformatics 2019-2020

Thomas Hamelryck & Jan Gorodkin

Deadline: Friday 24/01, 23h59

1 Protein part (2/3 of points)

1.1 Theory part (1/2 of points of protein part)

- In 2018, two revolutionary protein structure prediction methods appeared, namely DeepMind's **AlphaFold** and AlQuaraishi's **end-to-end prediction**. Your task is to write a comparison of these two methods by answering the following questions.
  - How do these methods deal differently with local versus non-local structure?
  - How is a "reference state" used by AlphaFold and why? What is the purpose of the reference state? Why does the end-to-end prediction not use a reference state?
  - To what extend are they knowledge-based and to what extend are they physics-based?
  - What are their major similarities and major differences?
  - In your opinion, what are their strengths and limitations compared to each other?
- Make use of the following sources:
  - The AlphaFold abstract at CASP13, 2018 (see Absalon).
  - The AlphaFold article at Proteins, 2019 (see Absalon)
  - AlQuraishi's end-to-end-prediction article, 2019 (see Absalon)
  - AlQuraishi's blog: "AlphaFold @ CASP13: What just happened?"
  - In addition, you can use the course's articles provided on Absalon, and any other reliable sources you see fit (provide references).
- Maximum 2 pages.

1.2 Practical part (1/2 of points of protein part)

Your task is to examine the variability in terms of RMSD of the side chains of the 18 amino acids excluding Gly and Ala. Gly and Ala are excluded because they lack degrees of freedom in their side chains due to their small size.

- As **protein data base**, use the top500 collection of high quality protein structures.
- Use Bio.PDB to implement the script.
  - Disregard any structures that cannot be parsed by the Bio.PDB parser, but ignore warnings.
- For each of the 18 amino acids (Gly and Ala excluded), select 1000 pairs randomly sampled from the protein data set with replacement.
  - The two amino acids in each pair should come from different proteins.
- Superimpose the side chain atoms using the optimal RMSD algorithm.
  - Side chain atoms are here defined as C-alpha, C-beta and anything attached beyond the C-beta. Main chain atoms (N, C and O) are excluded.

    - Exclude hydrogens.
  - Make a well-justified decision on how you are going to center the atoms before applying the optimal RMSD superposition.
  - Make a histogram of the RMSD distribution for each of the 18 amino acids.
    - Make sure all histograms use the same scale on the x- and y-axis.
  - Discuss and interpret the results.
- Tip: you can use Matplotlib to plot the histograms.
- You are allowed to make use of the exercise solutions provided by the course or your own exercise implementations.
- Maximum two pages, excluding figures.

2 RNA Part (1/3 of points)

The overall accuracy of RNA secondary structure prediction can be improved if for some stretches of the RNA sequence the structure is known. Knowledge about such substructures could come from either experimental or computational studies. Here, we will extend the Nussinov algorithm to consider a single known substructure at a given location. In addition, we require a minimum loop length of 3, such that $j > i + 3$ during the recursion.

1. Explain how this constraint can be implemented (hint: consider the initialization of the dynamic programming matrix). Then implement this constraint folding in your choice of Nussinov implementation (your own or one of those already available).
2. Use your implementation to predict the structure of the sequence:

   GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU

   for which the following substructure is already known (from position 26 to 42):

   .....(((((xxxxxxx))))).....

   See the nomenclature here, by clicking on "Show constraint folding".
3. Predict the structure for the full sequence (without the folding constraint); again set the minimal loop length to 3. Then, compute the base pair distance between the two dot-bracket strings you have obtained (with/without folding constraint). Discuss the difference in the structure and provide a sketch of the two structures. Does one of the structures resembles a known structure?
4. Annotate the RNA sequence by a method of your choice. Which structure prediction (constraint/unconstraint) is compatible with the annotation?
5. Run the RNAfold webserver without and with constraint. Compare the foldings and describe 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 2 pages).
2. The protein practical part (max 2 pages, excluding figures).
   (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 data set.

(d) A section with conclusions.

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:

1. An introduction to the Nussinov algorithm and its limitations in relation to the full energy folding of RNA secondary structure.
2. A materials and methods section that describes your Nussinov implementation and core aspect of the energy folding model.
3. The results of applying your implementation, energy folding and annotation of the given RNA sequence.
4. A section discussing the suitability (weaknesses and strengths) of the approach(es) to predict the RNA structure and your conclusions.
5. Maximum 3 pages, excluding figures.

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

3.3 General remarks

- Upload well-structured, well-documented code.
- Please provide references to the literature (or the occasional blog *if and only if* it's about referring to someone's opinion or to specific information not available elsewhere) – 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 effective automated method. Please do NOT exchange code or text. Please do NOT use verbatim quotes from sources without proper referencing.

Plagiarism WILL get you expelled from the university!