Computational biology – Home work

Master Computational and Mathematical Biology

Jacques van Helden

2019-12-19

Contents

Home work

During the practicals we retrieved different sequence sets.

- **A: CpG islands**
- **B: Random genome fragments** of the same sizes as the CpG islands
- **C: extended CpG islands** by adding 400 base pairs on each side of the CpG islands annotated in UCSC

We derived from this data

- an emission probability matrix whose parameters were estimated from the CpG islands
- an emission probability matrix whose parameters were estimated from the genomic background
- a transition probability matrix

In the following exercise, we will assume that the random genomic sequences contained a marginal proportion of CpG islands, and that we can thus use the emission probabilities of the genomic background as approximation for non-CpG islands.

We also *temporarily* assume that the UCSC annotations are perfect (i.e. they cover all the CpG islands and no other sequence). This assumption might however be reevaluated when interpreting the results, because annotations are never perfect.

The software implementation can be done in the language of your choice, but we expect to obtain a properly documented code that can run on another computer.

Choose **one** of the following problems. Each topic can be treated by a pair of students (ideally one biologist + one student from math/physics).

Problem 1: discrimination

1. Implement a program that computes the probability of each sequence given the CpG model, the genomic background, and assign a log-likelihood ratio score to each sequence of the datasets A and B.
2. Draw figures to depict the distribution of LLR scores in the respective sequence sets.
3. Compute the sensitivity (Sn), the False Positive Rate (FPR) and the positive predictive value (PPV) that would be obtained by assigning each sequence to the CpG group if it has a positive LLR, and to the other group otherwise.
4. Compute the same statistics if you would set the threshold to different values covering the range of observed LLR (you can even do so for all the observed values). Draw a curve with the Sn as a function of the FPR (ROC curve).

5. Develop your interpretation of the results (statistical performances, biological relevance, recommendations).

Problem 2: segmentation

1. Write a program that implements the Viterbi algorithm.
2. Use it to annotate the CpG islands in the dataset C.
3. Compare the annotations obtained with your Viterbi algorithm and those of the UCSC database, with a confusion matrix where
   - the columns correspond the reference sequence types (CpG or flanks in the UCSC annotations),
   - the rows correspond to the predicted sequence types (annotated by your Vitterbi algorithm),
   - the cells indicate the number of residues of each reference type assigned to each sequence type.
4. From this confusion matrix, derive the following statistics: sensitivity (Sn), False Positive Rate (FPR) and Positive Predictive Value (PPV).
5. Develop your interpretation of the results (statistical performances, biological relevance, recommendations).