Python exercises for the lecture
“Models and Algorithms in Bioinformatics”

Organizers: Peter Meinicke, Sina Garazhian
Submission date: 07.12.2024

Sheet 2: Hidden Markov Model (HMM) for G+C content analysis

a) Inspect and understand the script for DNA sequence simulation (`gc_simulate.py`). Generate a sequence of length 100000 as specified in the script and show histogram plots for the length distribution of the resulting foreground and background regions. What can you do to change the average length of foreground regions? For the following tasks use the generated test sequence (`seq_unlabeled`) and the corresponding state labeling of the generative processs (lower/upper case letters in `seq_labeled`) according to the specified random seed.

b) Implement the **Viterbi algorithm** for the HMM to analyze the test sequence. Use equal initial and final state transition probabilities (0.5) from/to begin and end states. Note that you must provide a numerically stable implementation that also works well for longer sequences!. Use the same parameter/model definition and naming scheme as in the simulation script. Make histogram plots for the length distribution of the predicted foreground and background regions. Compare the predicted state with the true region label of the generative process. Print the first few thousand labels of both state sequences for a ”visual” one-to-one comparison. Can you characterize the observable differences in terms of typical errors which show the limits of the path reconstruction? Calculate the *per-position* **sensitivity** and **specificity** (positive predictive value) with respect to the prediction of the label!

c) Change the foreground (”island”) emission probabilities in the simulation script according to a GC-content of 60% with equal probabilities for G and C as well as for A and T. Generate a new test sequence with same length as the one before. Repeat the analysis steps in b) for this sequence. Characterize the differences in the results for the two test sequences!

d) So far, we assumed equal prior state probabilities in the context of the begin state transitions. Estimate the prior state probabilities from the overall frequencies of generated states. Then, predict these probabilities without simulation just from the specified model parameters using the *steady-state* concept for Markov chains. Does the inclusion of initial steady-state probabilities change the predicted label sequence in b) ?

e) What could be done theoretically to investigate the uncertainty of the model at certain points in the sequence? Sketch an algorithm (without implementation) for marking sequence positions where the model may be unreliable with respect to the reconstruction of the states. Can you imagine how to combine Viterbi and *posterior* state probabilities for this purpose?

f) What modifications and extensions would be required to detect regions with anomalous GC-content in bacterial genomes in general? What kind of biological questions may be answered with such an approach? Take a look at the corresponding literature!

Additional literature

Soares, Siomar de Castro, et al.:
Genomic Islands: an overview of current software tools and future improvements
https://www.degruyter.com/document/doi/10.1515/jib-2016-301/pdf

Submission instructions

Please submit your solution by **7th December** using the corresponding homework folder in StudIP. Upload a **zip** compressed file containing all documents and data that you use in your solution. The file name should indicate the author, for example *Peter_Meinicke_Sheet2.zip*.
With your solution you should provide the Python source code together with the documentation within a **Jupyter Notebook** answering all the questions above. Use Markdown cells for your answers and explanations well separated from the Python code.
For the programming in Python/numpy do not use any code or functions from special bioinformatics libraries or toolboxes! Make sure that your code is readable and understandable for others, i.e. use descriptive names for the variable definitions according to the conventions on **Sheet 0**, do not use implicit variables (hard-coded values without a variable) and make use of comments!