02-425/725 HW2

February 16, 2024

1 Emission Probabilities

Consider an organism with genome consisting of "encoding regions" and "non encoding regions". Nucleotides $\{A, C, G, T\}$ are emitted with different probabilities in the coding and non-coding regions.

Use the attachment samples `encoding_regions.txt` and `non_encoding_regions.txt` to estimate the emission probabilities in each of these two states. Assume being in "encoding" and "non-encoding" only depends on nucleotide frequency and ignore all the other factors (start/stop codons, ribosome binding).

2 Gene-finding

Use your estimations of the emission probabilities in the previous problem to predict whether each of the sequences in `unlabeled.txt` belong to a coding or non-encoding region. Describe your method, and write a script.

Your solution, `HW2_2.txt` should have 10000 rows, each row either a 0, corresponding to non-coding regions, or a 1, corresponding to coding regions.

3 Building Hidden Markov Models

Use your estimations of emission probabilities from problem 1 to develop a hidden markov model to predict all the "encoding regions" and "non encoding regions" in the attachment `genome.txt`.

For transmission probabilities, try 0.001 for switching between, and 0.999 for staying in each state. At initial step, assume we can be in each state with probability 0.5.

Use HMM, not variable duration HMM. Assume being in "encoding" and "non-encoding" only depends on nucleotide frequency and ignore all the other factors (start/stop codons, ribosome binding).

Your solution, `HW2_3.txt`, should have one row with 1000000 digits, 0 for non-encoding regions and 1 for encoding regions.

4 Learning HMMs

Use your predicted regions in problem 3 to update transmission and emission matrices. Iterate between predicting hidden states and updating transmission/emission probabilities five times. How the matrices are changing? are they converging?

5 HMM Duration Distributions

Use your predicted regions in problem 3 to estimate the distribution of durations for encoding regions and non-encoding regions. Assume an exponential distribution

$$f(d, \lambda) = \lambda e^{-\lambda x} \text{ for } x \geq 0$$

and estimate λ in each region.