CSE182 Lecture 8,9 questions

Vineet Bafna

October 23, 2006

The questions are open ended, but should help you understand lectures better. Do these questions make sense? Are they helpful in following the lecture? Constructive feedback is appreciated.

1. Recall that we use a geometric method to distinguish coding versus non-coding. (L8, Slides 11,12). This method relies upon computing the angles α, β between two vectors. How do we compute the angles between two vectors?
2. Suppose you find a candidate sequence for which the E-score given by $\frac{\alpha}{\alpha+\beta}$ is 0.4. How can you increase your confidence in your prediction?
3. Define the problem of computational gene finding? What is a transcript? What is UTR? What are donor and acceptor sites?
4. Suppose you are given the entire cDNA transcript of a human gene. You are also told that the first intron contains a special regulatory site. Describe an algorithm to get the first intron assuming that the donor site begins with GT, and acceptor site ends with AG.
5. Suggest an algorithm that will return the UTR, given a 3' EST?
6. Recall the Viterbi algorithm for gene finding (L8, slide 25,26). How will you use the score computation to actually get gene coordinates? To answer this, you will need to create additional 'backtracking' matrices that store your choices, and use these to reconstruct the gene coordinates.
7. Suppose you 'knew' that the length of an exon was at least 40bp. Modify the equations in L8, slide 26 to incorporate this fact.
8. Suppose you 'knew' that a gene could have no more than 20 exons. Modify the equations in L8, slide 26 to incorporate this fact.
9. Vertebrate genomic sequence is generally depleted in the dinucleotide CG. This is because the C gets methylated, and may subsequently get converted to Thymine. However, ustream of expressed genes, the CG sequence is not methylated, resulting in a higher than expected frequency of the CG dinucleotide. Devise an HMM for detecting CG islands.
10. Given an alignment of donor sites as in L9, Slide 12, what is the best decomposition of the sequences into two subsets so that the dependencies are captured? What is the probability that $CGTA$ is a donor site?
11. Suppose we are given a gapless profile of length l that models an acceptor site signal. We are also told that immediately prior to the profile is a poly-pyrimidine track (sequence rich in C, T). Model the combined Acceptor site signal with an HMM.