BIOINFORMATIQUE

Examen Oral - Session de janvier 2008

Liste de questions

NB: For each algorithm, you should be able to analyze how its execution time and memory usage grow with the size of the dataset they are applied to. From this analysis, you should then be able to conclude about the type of problems to which the algorithm can be applied in practice.

1. Define the notions of multinomial and Markov chain models for nucleotide and amino-acid sequences. Describe how these models may be used for analyzing and comparing genomes.

2. Explain the notion of genomic signature and how it can be used to characterize and compare species using the chaos game representation of sequence content.

3. Explain the notion of statistical hypothesis test, type I and type II errors, power. Put this in contrast with decision theory and its derivation of optimal decision strategies minimizing the expectation of a cost. Explain two different applications of hypothesis tests in bioinformatics.

4. Explain the different uses of sequence alignments in bioinformatics. Describe the exact global and local alignment algorithms based on dynamic programming, and explain why more sophisticated but heuristic methods like BLAST are needed.

5. Explain what Hidden Markov Models are all about, why these are important in bioinformatics, and how they can be used. Explain the Viterbi and Forward-backward algorithms and their possible uses.

6. Explain the different notions of directed and undirected graph and tree. What are the possible uses of phylogenetic trees ? Explain the neighbor joining algorithm and its main properties.

7. Describe the uses of micro-array datasets. Explain the difference between supervised and unsupervised learning (give one example of each category) and their applications to the analysis of micro-array datasets.

8. What are the main similarities and differences between micro-array, proteomic and genotyping datasets, from the viewpoint of the information they contain and from the viewpoint of the methods needed to extract knowledge from them ?

9. What is the purpose of genetic association studies ? What kind of datasets are used for these studies ? What are the main statistical and algorithmic questions which are behind such studies ? Explain what linkage equilibrium is, where it comes from, and why it is useful ? Why is a 'complex' disease 'complex' to analyze ?