Bioinformatics – Lecture Notes

Announcements

Seminar: Modeling Spatial Dependencies for Data Mining
Dr. Wei-Li Wu
Wednesday, February 13
11 am – 12 noon
HT 1.303

Textbook web page http://www.cs.bc.edu/~clote/ComputationalMolecularBiology/
Has several programs of algorithms presented in the text.

Class 9 – February 12, 2002

1. Similarity Methods –
   a) Maximize the similarity between two sequences rather than minimizing the distance.
   b) Use similarity score function s(x,y) to compare characters x and y with $s(x,y) > 0$ for $x = y$ and $s(x,y) < 0$ for $x \neq y$. The score for a gap can be different than the score for a mismatch.
   c) transition = a substitution of a purine for a purine or a pyrmidine for a pyrimidine
   d) tranversion = a substitution of a pyr for a pur or a pur for a pyr
   e) Once again use dynamic programming algorithm.
   f) This can also be used for local alignments if the alignment for a random sequence is negative (which is guaranteed by the PAM matrices).

2. Multiple Sequence Alignment – This is used to determine the optimal aligment of several sequences. The text presents five methods (some quite briefly), namely Dynamic Programming, Gibbs Sampler, Maximum-Weight Trace, Hidden Markov Models, and Steiner Sequences.

Assume we are aligning k sequences with maximum length n.

   a) Dynamic Programming – Same as before except the distance matrix is $n \times n \times \ldots \times n$ (k times) and the weight function compares k letters.

   b) Gibbs Sampler – The Gibbs Sampler is based on the Gibbs Distribution (pages 47-52). Basically, perform local alignments on windows (subsequences) of fixed size. The optimal alignment is found after varying the window size and repeating.

c) Maximum-Weight Trace – Addresses a subclass of multiple sequence alignment problems, namely, the *sum-of-pairs* multiple sequence alignment problem. This method uses the creation of alignment graphs.

d) Hidden Markov Models – View the sequences to be aligned as a training set of observations that differ from an ancestor sequence as the result of a stochastic processes. The stochastic model that can best account for the sequences in the training set is determined. This method uses maximum likelihood and expectation minimization. This alignment is determined by how the sequences would match up with the ancestor sequence.

e) Steiner Sequences – This is related to multiple sequence alignment. In a method of DNA sequencing called *single-molecule DNA sequencing*, a single stranded DNA molecule is cut a single base pair at a time. The freed base flows down a glass tube by an optical sensor that determines the base. This technique has errors especially near the ends of the DNA. If this method is repeated many times, many copies of erroneous DNA are generated. This method computes an alignment of these sequences to find the actual sequence.