

Contents of the Tutorial

July 16, 2000

The goals of the tutorial are:

- to introduce attendees to the "computational genomics" world, by presenting, from the sequence determination to some sophisticated functional inferences, the different kinds of algorithms and models that are concerned and the main public data resources.
- to present to them some fundamental concepts of computational sequence analysis: important algorithms and models like "Dynamic Programming" or "Hidden Markov Models" will be explained.

In order to address in reasonable depth some algorithms and models, the tutorial will focus on one theme, "biological databases and search for similarities". "Biological databases" will include primary sequence data, and derived data like "motifs" or "domains". Since pairwise sequence comparison is at the heart of the methods and models involved in other problems like database similarity searching, and multiple alignment, it will be reviewed in the first part of the tutorial. The second part will address the problem of sequences database similarity searching, using the Fasta and Blast programs. Then, the third part will introduce to multiple alignments and their use through abstract representations like regular expressions, profiles and Hidden Markov Models (HMM).

Introduction

- computational genomics and use of sequence comparison;
- search for similarities in biological databases;
- biological databases;

1 Pairwise sequence comparison

1.1 Algorithms

1.1.1 Basic sequence comparison: dot-plots

1.1.2 Alignments

- the alignment as a path in a graph;
- score of an alignment, similarity vs distance measures;
- Dynamic Programming (Needleman & Wunsch and Smith & Waterman algorithms).

1.1.3 Scoring models

- Protein similarity matrices
- Matrices for nucleic acids

2 Fasta & Blast programs

2.1 Algorithm of Fasta

2.2 Algorithm of Blast1

2.3 Blast2 programs

2.4 Statistics

3 Motifs and Multiple alignments

3.1 How to compute a multiple alignment?

3.1.1 Global multiple alignment

- Simultaneous alignment by Dynamic programming (e.g. MSA, DCA)
- Progressive alignment (e.g. ClustalW)
- Pattern induced global alignment (e.g. PIMA, DIALIGN2)

3.1.2 Local multiple alignment (motif inference)

- Gibbs and MEME: stochastic optimization
- Pratt
- Consensus

3.2 Representation of the information contained in a multiple alignment

3.2.1 Consensus

3.2.2 Regular Expressions

3.2.3 Profiles

3.2.4 Hidden Markov Models

3.3 Examples

3.3.1 Available data resources

Databases like Prosite, Prints, Blocks, etc.

3.3.2 Use of these concepts in the context of a sequences DB search

Examples of PSI-blast and PHI-blast.