

# ISMB'02 Tutorial

## Introduction to Computational Sequence Analysis

Frédérique Galisson

### Motivation and goals

Most of the new data entering the biological databases now come from whole genome sequencing projects. From the genes themselves to the structural or functional properties of the predicted proteins, most of the biological features discovered through genome sequencing projects are inferred by computational analysis of the sequences. All these results have completely modified the world biologists are working in, offering them new data and tools, but with few clues for enabling them to fully exploit these. In order to evaluate the biological relevance of the available methods and data and to make the relevant choices with respect to particular biological problems, an understanding of what is inside the programs and databases is needed. The computational representation and analysis of biological sequence data are based on theoretical (mathematical or biological) models and make use of algorithms leading to methods whose efficiency, accuracy, sensitivity and specificity may vary greatly from one program to another and depend on their parameters.

In this context, the goals of the tutorial are:

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

### Intended audience

This tutorial is primarily aimed at biologists who do not have any real background in bioinformatics, and who wish to understand the models and meth-

ods underlying the main approaches used in computational sequence analysis. It has already been given at ISMB'00, and from this experience (and the feedback from the attendees), it may as well suit computer scientists.

## **Contents and tentative outline**

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".

### **1 Pairwise sequence comparison**

- Algorithms: Dot-plots, Alignments (score of an alignment, similarity versus distance measures, Dynamic Programming procedures)
- Scoring models: Modeling of gap weights, protein similarity matrices, matrices for nucleic acids.

### **2 Looking for similarities in sequence databases**

- Sequence Databases: nucleic vs proteic, general and exhaustive vs specialized, etc
- The Fasta and Blast programs
- Scores, statistics and significance of a search.

### **3 Motifs and Multiple alignments**

- algorithms used for building multiple alignments.
- representation of the information extracted from several aligned sequences: consensus, regular expressions, profiles, HMMs.
- available data resources: databases like prosite, blocks, pfam... and programs.
- use of these concepts in the context of a sequence DB search: examples of PSI-blast and PHI-blast.

## Tutor

Frédérique Galisson, PhD, is currently working at the University of Lausanne (Switzerland) and the Swiss Institute of Bioinformatics. She is involved in several bioinformatics teaching activities: in charge of a 35 hours course to researchers at the Faculty of Medicine, in charge of a 42 hours course (biology and bioinformatics) to computer science students at the EPFL (*“École Polytechnique Fédérale de Lausanne”*), involved in the Master degree in Bioinformatics (Universities of Geneva and Lausanne), and contributing to other courses like the EMBnet courses.

Previous Bioinformatics teaching experience includes courses that she developed and taught (50 hours, two times a year) for five years (1996-2001) to biology researchers at the Pasteur Institute (Paris), as well as several one-week (25-30 hours) courses at other places: Cornell/Rockefeller/Sloan-Kettering tri-institutional MD-PhD program (MD-PhD and graduate students), New-York city, January 2001 and 2002; Pasteur Institute of Cambodia (students and researchers, biology and medicine), Phnom-Penh, September 2001; University of the Western Cape, South-Africa (honors students, biochemistry), February 2000; IRD Montpellier, France (PhD students and researchers, biology), May 2000.

This tutorial, which has been built from the courses mentioned above, has already been given at ISMB'00, San-Diego, August 2000, where it received good evaluations from the participants.