

ISMB'03 Tutorial

The EM algorithm and some of its applications in molecular biology

Frédérique Galisson

Motivations and objectives

When analysing biological data like sequences, many questions can be formalized as the clustering of the data into several groups, corresponding to different biological features. When these features can be represented *via* statistical models (like statistical profiles for modeling sequence motifs, or Markov chains for coding regions), mathematical tools are available, enabling one to compute the probability of any model given any observed data. If the models which are expected to be found in the data are known (supervised clustering), it is then easy to score each model with respect to the data, and classify the data according to these scores. When the model parameters are not known, and have to be guessed from the data (unsupervised clustering), the problem can be solved using the *Expectation-Maximization* (EM) algorithm. Developed in 1977 by Dempster et al., it can be used for a great variety of problems with incomplete or missing data. It has been applied to several computational biology problems like the sequence motif inference problem, the unsupervised search for coding regions in sequences, or more recently, to the clustering of microarray expression data.

The objective of this tutorial is to give a unified presentation of the application of EM to these different problems, by first explaining a form of EM which can be used for clustering problems (EM for mixture models), and then showing through three examples (motif inference, search for coding regions, and microarray data clustering) how different problems can be formalized the same way, with mixtures of statistical models specific to each problem, used in combination with the same EM algorithm.

Intended audience

The tutorial is aimed at researchers in computational biology, interested in the theoretical aspects of the computational analysis of biological data. Some background is expected: elements of mathematical formalism, basic knowledge in statistics and in molecular biology.

Contents and tentative outline

1 The EM algorithm and Mixture models

- **Theoretical background: Bayesian inference**
- **Maximum Likelihood estimation of model parameters**
 - simple case with complete data
 - missing data: the "mixture" case
 - supervised versus unsupervised learning
- **The EM algorithm for mixture models**
 - Formal mathematical presentation of the algorithm
 - Mixture models
 - EM for Mixture models: detail of the E and M steps

2 Motif inference

- **A two-component mixture model**
 - Modeling of the motif: profiles and HMM
 - Modeling the "background" sequences
- **Supervised learning: scanning a profile along a sequence**
- **EM with such a model** (Lawrence and Reilly, 1990)
- **The MEME program** (Bailey and Elkan, 1995)
- **Other examples**

3 Search for coding regions

- **Models of coding regions**
 - Statistical biases of coding sequences
 - Inhomogeneous Markov chains of period 3

- A seven component model accounting for 6 reading frames and the non-coding hypothesis.

- **Supervised learning and the example of Genemark**

- **Unsupervised search for coding regions with EM**

- Theory: EM with this model
- Previous related work (Audic & Claverie, 1998 ; Hayes and Borodovski, 1998 ; Viari 2001, unpublished)
- Implementation and validation (F. Galisson et al.)

4 Clustering of gene expression data

- **Microarray data and the problem of their clustering**

- **Model-based clustering for gene expression data** (Yeung et al, 2001 ; McLachlan et al, 2002)

- **Other examples of applications of EM or other bayesian techniques to the analysis of microarray data**

Tutor

After a PhD in Molecular Biology, and a post-doc in Genomics, FG started working in the Bioinformatics field in 1996, in the Scientific Computing group of the Pasteur Institute in Paris (France). Since 2001, she has been working at the University of Lausanne (Switzerland) and the Swiss Institute of Bioinformatics.

For the last seven years, she has been teaching computational biology, mainly to biologists, but also to computer scientists, at levels ranging from undergraduate students to confirmed researchers.

She has been working on the application of the EM algorithm to the search for coding regions: a general version of EM for mixture models has been implemented, as well as probabilistic models for coding regions. The implementation has been validated on synthetic data, and the method also proved to work on bacterial genomes (unpublished).

Previous ISMB experience: she taught another tutorial, "Introduction to computational sequence analysis", at ISMB in 2000 and 2002. She was in charge of the organization of the ISMB tutorial session in 1998, 2001, and 2002.