HIDDEN MARKOV MODELS AND MULTIPLE ALIGNMENTS OF PROTEIN SEQUENCES

Pavle Goldstein  
*Department of Mathematics, University of Zagreb*  
Pavle.Goldstein@math.hr

Maja Karaga  
*Department of Mathematics, University of Zagreb*  
karaga@math.hr

Mate Kosor  
*Department of Mathematics, University of Zagreb*

Ivana Nižetić  
*Department of Mathematics, University of Zagreb*

Marija Tadić  
*Department of Mathematics, University of Zagreb*

Domagoj Vlah  
*Department of Mathematics, University of Zagreb*

**Abstract**  
A multiple alignment algorithm for protein sequences is considered. Alignment is obtained from a hidden Markov model of the family, which is built using simulated annealing variant of the EM algorithm. Several methods for obtaining the optimal model/alignment are discussed and applied to a family of globins.

**Keywords:** multiple alignment, hidden Markov model, simulated annealing, expectation maximization, suboptimal alignment.
1. Introduction

Multiple alignments of protein sequences are one of the most important computational methods for studying proteins. Multiple alignment of a protein family can be used to describe the evolutionary relationships within the family and build a phylogenetic tree, to detect structurally important or functional sites, and infer structure and function. Let us also point out that, given sufficient similarity between the sequences, “the correct” multiple alignment of sequences will actually describe the best possible match of associated three-dimensional objects (i.e., proteins in their folded state). Consequently, MSA algorithms, or, given the uncertainties involved, MSA strategies, are a topic of great interest in computational molecular biology.

In this note, we describe several MSA strategies based on hidden Markov models. Our aim was to design a robust procedure with resulting alignments comparable in quality to the ones produced by heuristic algorithms, primarily CLUSTALW (cf. [4]). It should be pointed out that, since they were introduced into computational biology approximately fifteen years ago (cf. [1]), HMMs found various applications, primarily as family profiles, but also as MSA tools (cf. [6]). However, those methods, in contrast to CLUSTALW, require a representative sample of the family to work with. Here we show that, when working with an unrepresentative sample, the problem of finding the optimal alignment is quite different from the problem of determining the optimal model. We also propose modifications to the model that we believe will rectify this problem.

Let us briefly describe our setup and introduce some notation: in the next Section, we give a short description of the HMM used. Sections 3 and 4 contain details on expectation maximization implementation and suboptimal alignment procedures, respectively, while in the last Section we show the results of our tests. Throughout the paper, the family of protein sequence will be denoted by $x = \{x^i\}^n_{i=1}$, where each $x^i = x_1^ix_2^i \ldots x_k^i$ is a finite word in the alphabet $\mathcal{A}$ of 20 standard amino acids. For a family $x$, multiple alignment of $x = \{x^i\}^n_{i=1}$ is the family $\text{MA}(x) = \hat{x} = \{\hat{x}^i\}^n_{i=1}$, where each $\hat{x}^i$ is a word in the alphabet $\mathcal{A} \bigcup \{-\}$, and $\hat{x}^i$ restricted to $\mathcal{A}$ equals $x^i$. Since we do not allow columns consisting entirely of gaps, for each $l$, there exists $j$ such that $\hat{x}^j_l \neq -$.

2. Hidden Markov Models

The putative Markov process under consideration consists of the emission of a single amino acid. It is common to use the HMM from Figure 1 to model this process (cf. [1, 2]). Here, the squares $M$ correspond to the match states, diamonds $I$ to the insert states and circles $D$ to the silent delete states. The model is determined in terms of parameters – emission probabilities $e_S(b)$, probability of state $S$ emitting the symbol $b \in \mathcal{A}$, for $S$ being $M$ or $I$, and transition probabilities $a_{ST}$. 