A Jumping Profile HMM for Remote Protein Homology Detection

gobics.de
Göttingen Bioinformatics

Anne-Kathrin Schultz and Mario Stanke
Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Universität Göttingen, Germany
contact: {aschult2, mstanke}@gwdg.de

Abstract

We address the problem of finding new members of a given protein family in a database of protein sequences. Given a multiple sequence alignment (MSA) of the sequences in the protein family, we would like to score each candidate sequence in the database with respect to how likely it is that it belongs to the family. Successful methods for this task are profile Hidden Markov Models (HMM), like HMMER [Eddy, 1998] and SAM [Hughey and Krogh, 1996], and a so-called jumping alignment (JALI) [Spang *et al.*, 2002].

We developed a Hidden Markov Model which can be regarded as a generalization of these two methods: At each position the candidate sequence is either aligned to the whole column of the MSA or to a certain reference sequence. Thus our model catches both the horizontal and vertical information in the MSA. It is called a **jumping profile HMM**.

Successful Known Methods: Profile HMMs and JALI

A profile HMM uses mainly a summary of the columns of the MSA to align a candidate sequence to the MSA and to define a probability that measures how well the candidate protein fits into the protein family. It contains information about conserved key residues shared by most or all sequences of the family. The penalty for a deletion or an insertion in the candidate sequence depends on its position in the alignment to the MSA.

JALI is a jumping alignment of the candidate sequence to the MSA of the protein family. At each position the candidate sequence is aligned to one reference sequence of the MSA using a scoring matrix. This reference sequence can change within the alignment. Such a change of the reference sequence is called a *jump* and is penalized by subtracting a constant jump cost from the score. The jumping alignment method can catch dependencies between residues at different positions of the sequence and appears to be more robust when the MSA of the family is partially wrong.

Our Generalization: Jumping Profile HMM

We are given a MSA of k rows and a candidate sequence. At each position the candidate sequence is either aligned to the whole column of the MSA or to a certain reference sequence: We say that we are in the *column mode* or in a *row mode* of the HMM.

- **Column mode**: (red part of Figure 1)
  As in a profile HMM each consensus column of the MSA is modeled by three states: match (**M**), insert (**I**) and delete (**D**).Match states model the distribution of residues in this column, they emit the amino acids with a probability which depends on all residues in this column.
- **Row modes**: (first k rows of Figure 1)
  Each sequence of the alignment corresponds to a row mode, i.e. for each amino acid in the alignment we have three states: match (**M**), insert (**I**) and delete (**D**). The emission probabilities of the amino acids in the match states only depend on the residue found at this specific position in the alignment.

Insert and delete states allow for insertion or deletion of one or more residues at a certain position in the alignment. The states are connected by transitions to which transition probabilities are assigned.
To estimate the emission and transition probabilities we use Dirichlet mixtures [Sjölander *et al.*, 1996], [Wistrand and Sonnhammer, 2004] for all states.
Changes between the column mode and row modes and changes between different row modes are possible and penalized. They allow to align the candidate sequence partially to the columns of the MSA and partially to a reference sequence, which may change within the alignment.
These transitions are called jumps and are shown just for three example states: $M_{1,2}$ (cyan), I_{1,n_1-1} (green) and D_{1,n_1-1} (blue) (Figure 1).

Aligning a sequence to the MSA with respect to the jumping profile HMM corresponds to finding the most probable path through the model that emits the sequence.

A jumping profile HMM is a generalization of profile HMMs and jumping alignments in the following sense. If we only take the column mode of our model we get a profile HMM. Leaving out the column mode our model is a probabilistic analogon of JALI. Jumps in JALI correspond to transitions between the states of the different row modes.

Jumping Profile HMM - Features

- The model catches both the horizontal and vertical information in the MSA. It considers information about conserved columns of the alignment as well as information about conserved sequence patterns common to only some of the sequences.
- It can both find sequences which are similar to the profile of a protein family and sequences which are only partially similar to one sequence of the family.
- It automatically includes a pairwise comparison to each sequence of the family. Thus it is less impaired by a bad or wrong MSA as a profile HMM.

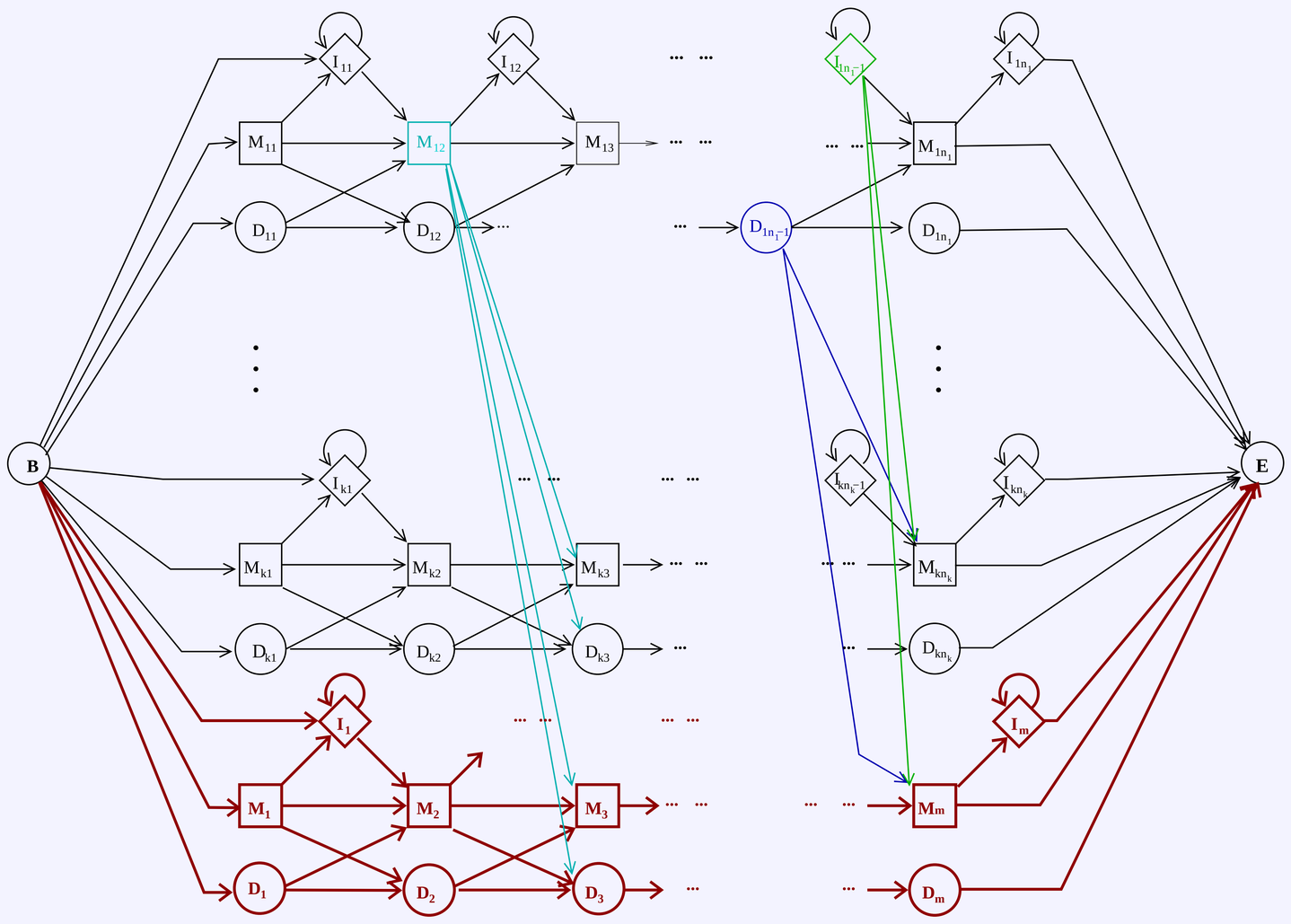

FIGURE 1: Architecture of a jumping profile Hidden Markov Model

References

[Eddy, 1998] Sean R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[Hughey and Krogh, 1996] Richard Hughey and Anders Krogh. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *Comput. Appl. Biosci.*, 12(2):95–107, 1996.

[Sjölander *et al.*, 1996] Kimmen Sjölander, Kevin Karplus, Michael Brown, Richard Hughey, Anders Krogh, I. Saira Mian, and David Haussler. Dirichlet Mixtures: A Method for Improved Detection of Weak but Significant Protein Sequence Homology. *Comput. Applic. Biosci.*, 12:327–345, 1996.

[Spang *et al.*, 2002] Rainer Spang, Marc Rehmsmeier, and Jens Stoye. A Novel Approach to Remote Homology Detection: Jumping Alignments. *J. Comp. Biol.*, 9(5):747–760, 2002.

[Wistrand and Sonnhammer, 2004] Markus Wistrand and Erik L.L. Sonnhammer. Transition Priors for Protein Hidden Markov Models: An Empirical Study towards Maximum Discrimination. *J. Comp. Biol.*, 11(1):181–193, 2004.