Hidden Markov Models in Protein Modeling

Marco V. Bayas November 16, 2001

Abstract.

The use of Hidden Markov Models (HMM) in protein modeling is described. Sequence alignment based on profile HMMs can help identifying protein family members and present some advantages. This possibility is discussed.

Introduction.

The functional and structural characterization of new proteins can be done by taking advantage of their evolutionary relation with proteins of known structure or function. Statistically significant identification of homologous proteins is crucial for this.

One way to look for homologous sequences is the use of statistical profiles of protein families. A profile is a model that shows the amino acid distribution for each position in the family. A systematic method for constructing profile models is provided by Hidden Markov Models.

Profile Hidden Markov Models

Use of HMMs for representing profiles of multiple sequence alignments was first used by Krogh et al [3]. They introduced an HMM architecture consisting of three sets of states. “Match” states describe the conserved structure in a protein family. Additionally, “insert” and “delete” states allow for insertion or deletion of one or more residues respectively. All transitions probabilities between states as well as all character emissions in the insert and match states are fixed based on the information of a protein family. Under these circumstances, emission of amino acids as the position moves from the start to the end node in the model generates protein sequences. The probability of any sequence is computed by multiplying the emission and transition probabilities along the path. A diagram of a small profile HMM is included in appendix 1.

The different probabilities of the model can be set in two ways. An HMM can be trained from initially unaligned sequences of an identified family. The number of sequences considered is important for the significance of the model, for example, in the case of the globin family 200 sequences is enough [3]. Alternatively, an HMM can be built from prealigned sequences. The last one is the relatively easiest way. Once an HMM is available, regardless of its complexity, the same standard dynamic programming algorithms can be used for aligning and scoring sequences with the model, making it possible to discriminate between family and non-family members.

In general, the layout of a model depends on the specific application. For example, for prediction of transmembrane protein topology (see appendix 2) [4] the layout includes submodels designed to model specific region of a membrane protein, such as the transmembrane helix core. These submodels contain several HMM states in order to model the lengths of the various regions. Transitions between submodels make sure that the constraints associated to helical transmembrane proteins are hold. This model predicts 97-98% of transmembrane helices.

Advantages of Profile HMMs

Profile HMMs differ from the more conventional techniques based on pairwise alignments. The alignments generated by these methods are strongly dependent of the particular values of parameters required by the model, in particular the gap penalties. In a profile HMM, the gap costs are not arbitrary numbers. This is because the transition probabilities involving insertion and matching are correlated. In fact, the sum of the probabilities of all of the possible transitions from one state must be equal to one.

Additionally, profile HMMs implicitly includes a cost for inserted residues whereas in traditional alignment inserted residues have no cost besides the affine gap penalty. Not including a cost for insertions would mean that unconserved insertions in protein structures have the same residue distribution as proteins in general which is not necessarily the case. In fact, insertions tend to be seen most often in surface loops of protein structures, and so have a bias towards hydrophilic residues [1]. Profile HMMs can capture this information in the insert state emission distribution, making alignments more realistic.

Discussion.

In order to get a profile HMM it is necessary to run a multiple alignment in the first place. To do that it is necessary to use alternative techniques for alignment. So, in principle, the same biases of the other techniques could be introduced in profile HMMs. However, this problem can be reduced by increasing the number of sequences considered. This reduces the applicability of profile HMM to protein families with a larger number of members.

A profile HMM makes it possible to consider explicitly some important biological facts such as bias toward specific kinds of amino acids or length of transmembrane domains. If the training process is adequate this facts will be reflected in the values of the emission and transition probabilities

The identification of homologous proteins is important not only for structural studies but also for phylogenetic studies. Provided the profile HMM of a family is known the identification of evolutionary distant related proteins is more realistic. This makes the method ideal for the construction of phylogenetic trees.

References

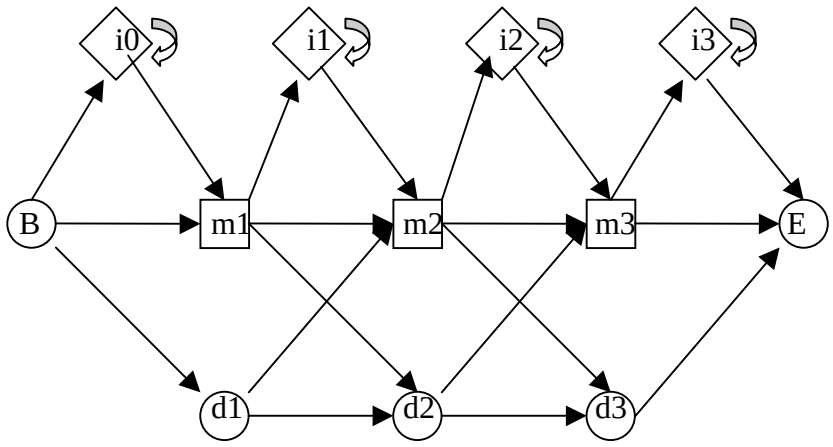
1. Eddy, S.R. 1998. *Profile Hidden Markov Models*. Bioinformatics. 14: 755-763

2. Karplus, K., Barret, C., Hughey, R. *Hidden Markov Models for detecting remote protein homologies*. Bioinformatics, 14: 846-856.

3. Krogh, A. et al. 1994. *Hidden Markov Models in Computational Biology: Applications to Protein Modeling*. J. Mol. Biol. 235: 1501-1531

4. Krogh, A. et al. 2001. *Predicting Transmembrane Protein Topology with a Hidden Markov Model: Applications to Complete Genomes*. J. Mol. Biol. 305: 567-580

Appendix 1. A small profile HMM with three consensus columns.



The three columns are modeled by three *match* states (m1, m2, m3) each of which has 20 residue emission probabilities. Additionally, there are four *insert* states (i0, i1, i2, i3) each of them also have 20 emission probabilities. Finally, there are three *delete* states (d1, d2, d3) without emission probabilities. The *begin* and *end* states define the extremes of the sequence. The arrows indicate the possible transitions between states. Each of the transitions has a specific probability.

Appendix 2. Layout of a HMM for transmembrane protein topology prediction (ref 4)