Making hidden Markov models more biologically realistic: Improvements in remote homolog detection and alignment quality

Melissa Cline and Kimmen Sjölander
joint work with
Christian Barrett, Marc Hansen, David Haussler, Richard Hughey, and Kevin Karplus

Hidden Markov models have been successfully applied to the problems of database searching and multiple sequence alignment of protein families and domains, and for finding coding regions in DNA. In previous work we have shown how an HMM can be constructed that identifies a set of positions that describe the (more or less) conserved first-order structure in a set of sequences. In biological terms, this corresponds to identifying the core elements of homologous molecules. The model also provides additional information, such as the probability of initiating an insertion at any position in the model and the probability of extending it. The structure of the model is similar to that of a profile, with position-specific insert and delete probabilities.

In our work with building HMMs for proteins, we have found that incorporating prior information in the form of mixtures of Dirichlet densities over typical amino acid distributions increases the generalization capacity of these models.

These techniques have been limited by the omission of biologically important information. In recent work, we address this deficit, and include biologically relevant information where possible. We have found that by including this information, we are able to refine existing alignments and improve results in database searches for remote homologs.

Much of this recent work has developed out of necessity. We have been building HMMs for protein domains—compact regions of proteins which fold (more or less) independently in solution—in an attempt to parse proteins of unknown structure into their constituent domains, and thus predict the overall fold of the protein. To do this well, we need to first refine the existing alignment upon which the model is based, and second, generalize the model we obtain to be able to identify other regions in proteins which are structurally similar, but may have very low primary sequence identity.

What makes this kind of generalization possible is that the alignments we are using as input to our model-building process contain structural information. In particular, a large fraction of the columns contain information regarding the particular secondary structure for the column (helix, strand, turn or other), and solvent accessibility. We use this information in several ways.

First, this information is used to control the insert and delete probabilities. For instance, insertions or deletions in helices are highly unlikely, whereas in loop regions these are more likely. A one-residue insertion in a beta strand might have some probability, but a two-residue insert is highly unlikely. The first is a beta bulge, but the second disrupts the structure. We use this information in a way that is analagous to our method for regularizing amino acid distributions in proteins, as priors over such probability distributions. In this case, the probability distribution involves transitions between states in the HMM.

We also use this information to develop structurally informed Dirichlet priors over the expected amino acids in each position. Rather than a single Dirichlet mixture density which is required to work well at every position, we allow each position to specify which density it will use to compute the expected amino acids at that position. This allows positions in buried strand environments to select a prior that is trained on alignment columns obtained from such environments, and positions which are exposed to select a prior that is trained on alignment columns from exposed positions.

We have also recently developed a method for identifying subfamilies in an alignment. Once these subfamilies are identified, we use this information to build subfamily models. These models share information for positions which correspond to the common structure underlying all the subfamilies, and focus on their own specific signals in the columns which differentiate the subfamilies. We also use the subfamily identification to guide the construction of weights for the sequences.

These methods, used jointly, result in models which have greater effectiveness at database discrimination and remote homolog detection.