# The Grammatical Structure of Protein Sequences in a Single Non-linear Hidden Markov Model; Prediction of Coding Regions, Secondary and Tertiary Structure

**Christopher Bystroff**
bystrc@rpi.edu
Deptartment of Biology, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA

**Keywords:** protein folding, hidden Markov model, structure prediction, gene finding, local structure

## 1 Introduction

We describe a hidden Markov model, HMMSTR, for general protein sequence based on the I-sites library of sequence-structure motifs [1]. Unlike the linear HMMs used to model individual protein families, HMMSTR has a highly branched topology and captures recurrent local features of protein sequences and structures that transcend protein family boundaries. The model extends the I-sites library by describing the adjacencies of different sequence-structure motifs as observed in the database, and achieves a great reduction in parameters by representing overlapping motifs in a much more compact form. The HMM attributes a considerably higher probability to coding sequence than does an equivalent dipeptide model, predicts secondary structure with an accuracy of 74.6% and backbone torsion angles better than any previously reported method, and predicts the structural context of beta strands and turns with an accuracy that should be useful for tertiary structure prediction.

## 2 Methods and Results

### Formulation of the hidden Markov model

A hidden Markov model is a network of Markov states connected by directed transitions. Each state emits symbols representing sequence and structure.

The non-redundant subset of all proteins of known structure has been encoded as a linear sequence of discrete amino acid and structural symbols, augmented by an amino acid frequency profile obtained by multiple alignments. The amino acid of the parent sequence is denoted by $O_t$, and the profile by $\{O_t^m\}$ $(1 \leq m \leq 20)$. For the structural identifiers the following nomenclature is used: 3-state secondary structure $D_t = \{H, E, \text{ or } L\}$, discrete backbone angle region $R_t = \{\text{one of 11 regions}\}$ and the context symbol $C_t = \{\text{one of 10 symbols}\}$. A sequence $S = s_1, s_2, \ldots, s_T$ is given by $s_t = \{O_t, \{O_t^m\}, D_t, R_t, C_t\}$ $(1 \leq t \leq T)$.

The utility of the HMM to model database sequences and structures is based on the notion of a *path*, which is a sequence of states through the HMM, denoted $Q = q_1 q_2 \cdots q_T$. Thus, the joint probability of a symbol sequence $S$ and a state sequence $Q$ given the HMM $\lambda$ is obtained by:

$$P(S, Q \mid \lambda) = \pi_{q_1} B_{q_1}(s_1) a_{q_1 q_2} B_{q_1}(s_2) \cdots a_{q_{T-1} q_T} B(s_T) \tag{1}$$

where $B_q(s)$ is the probability of observing symbols $s$ at state $q$, and $a_{ij}$ is the probability of a transition from state $i$ to state $j$, and $\pi_i$ is the probability of starting in state $i$.

This probability summed over all possible paths gives us a measure of the relative likelihood that the sequence represents a coding region. The path or paths of maximum probability give us a sequence of Markov states whose structural symbols $D$ and $R$ represent the predicted secondary or local structure. Similarly, the context symbols $C$ associated with each Markov state tell us, in combination with the secondary structure prediction, whether certain tertiary contacts are predicted to be made.

### Initialization based on the I-sites motif library and training on known structures

An I-sites motifs, 262 in number ranging in length from 3 to 15 residues, may be represented as strings of Markov states, or Markov chains, initialized and numbered sequentially. To combine them into a single model, the Markov states were merged using a measure of similarity that incorporates sequence and structure over a variable stretch of the chain. The order of merging was based on state-sequence similarity and a few contraints on the topology. The resulting HMM is a sparce, directed graph with many branches and cycles. A single non-emitting state was inserted between all sink states and all source states.

Expectation maximization was performed with a generalization of the algorithm described in Rabiner [2]. The training set consisted of a non-redundant set of 794 proteins of known structure. An independent test set of 73 proteins was set aside and used only in the final evaluations.

### Prediction of coding regions,local and secondary structure, and structural context.

The per-position information content of a protein sequence using HMMSTR, relative to a simple dipeptide frequency Markov model (almost equivalent to a fifth-order MM on nucleotides), was averaged over all of the proteins in the test set, using only a single parent sequence for each. A net increase in information content of 0.0188 nats/position was provided by the HMM. The dipeptide model in turn provided 0.0072 nats/position of information over the single-position "background" model. (Although the incorporation of profile data in the equation adds ten-fold to the information content, that additional information is not relavent to the gene finding problem since the presence of homologous sequences is itself a much stronger signal for coding regions.)

Three-state secondary structure predictions were made by calculating the single-position marginal probability of each secondary structure state over all possible Markov state paths. The results on the test set were 74.6% correct; comparable to the best methods available. Backbone angle prediction were made in a similar way, computing the marginals for each of 11 regions of Ramachandran angle space. Using the MDA measure [1], HMMSTR correctly predicted 59% of all eight-residue segments, compared to 48% for the I-sites library and 43% for the best secondary structure predictors. Context symbols were predicted for turn residues if they fell between two strands. Predictions indicated that the model can discriminate between hairpin and diverging beta-turns, and to some extent can predict whether a beta-strand is in the middle of a sheet or at the end.

The structure of the HMM for predicting protein local structure is suggestive of a model for grammatical structure in written language, with recurrent short patterns arranged into words, and words into phrases. By further incorporation of rules for tertiary structure formation, this method may eventually link the phrases into full sentences, addressing the *ab initio* folding problem.

## References

[1] Bystroff, C. andBaker, D., Prediction of local structure in proteins using a library of sequence-structure motifs, *J. Mol. Biol.*, 281(3):565-77, 1998.

[2] Rabiner, L.R., A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proc. IEEE*, 77(2):257–286, 1989.