Hidden Markov Model for protein secondary structure

Juliette Martin, Jean-Francois Gibrat, and Francois Rodolphe

Unité Mathématique Informatique et Génome,
INRA, Domaine de Vilvert,
78350 Jouy-en-Josas Cedex, France
(e-mail: [Juliette.Martin,Jean-Francois.Gibrat,
Francois.Rodolphe]@jouy.inra.fr)

Abstract. We address the problem of protein secondary structure prediction with Hidden Markov Models. A 21-state model is built using biological knowledge and statistical analysis of sequence motifs in regular secondary structures. Sequence family information is integrated *via* the combination of independent predictions of homologous sequences and a weighting scheme. Prediction accuracy with single sequences reaches 65.3% and raises to 72% of correct classification with profile information.
Keywords: α-helix, β-sheet, prediction.

1 Introduction

Proteins are the main actors of living cells. Many cellular constituents are made out of proteins. Almost all enzymes are proteins, cellular pumps and motors are made out of proteins.

The function of a protein strongly depends of its 3D-structure. For instance, enzymes need to have a tight spatial complementarity with their substrates (reaction partners). Thus knowledge of a protein structure gives relevant clues to its function.

Since genome sequencing started, the even widening gap between the number of protein sequences and protein structures available in databases enhances the utility of structure prediction methods. Because of the structure-function relationship, structures are more conserved than sequences during evolution and therefore different sequences can have the same 3D structure.

Structure prediction methods fall into two categories:

- comparative modeling if a related structure is known and can be used to derive a global model,
- *de novo* prediction if there is no related structure available.

We are presently interested in the latter. *De novo* prediction methods often require a first step of local structure prediction: secondary structure prediction in our case. Three canonical classes of secondary structures are considered : α-helices, β-strands and coil, see figure 1.

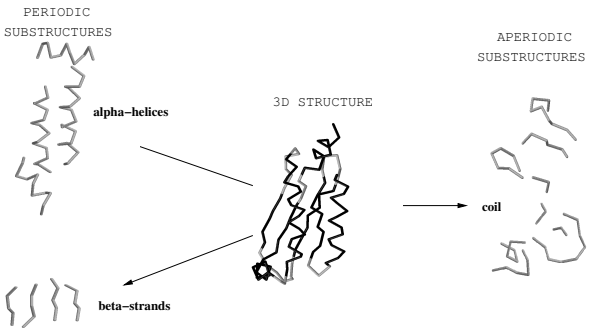

Fig. 1. Secondary structure of proteins. A 3D protein structure (center) can be described in term of secondary structures: α-helices, β-strands (left side) and coil (right side). Only C-α are shown, periodic substructures are indicated in black in the full 3D structure.

α-helices and β-strands are geometrically periodic sub-structures frequently occurring in 3D structures (about 50% of residues in proteins are involved in α-helices and β-strands). Coil denotes all sequence segments which do not fall into one of these two categories.

We use Hidden Markov Models to predict the three classes of secondary structure. The model is built using prior biological knowledge and pattern analysis in protein sequences.

2 Data set

The data set is a subset of 2530 structural domains taken from ASTRAL 1.65 [Brenner *et al.*, 2000], determined by X-ray, with a resolution factor less than 2.25 $\mathring{A}$ and less than 25% sequence identity. Secondary structure definition is given by an assignment method developed in our laboratory (manuscript in preparation) or by STRIDE method [Frishman and Argos, 1995]. 489743 residues have a defined secondary structure in our data set. 2024 sequences, randomly selected, are used in a four-fold cross validation procedure: three quarters of these sequences are used for parameter estimation and one quarter is used for the test. The remaining 506 sequences are used as an *independent* test set. This test set is never used to estimate model parameters. The use of an independent test set allows to check that no bias is introduced during the model design when searching for characteristic motifs in secondary structures (see hereafter). The number of residues with a defined secondary structure are 94790, 101521, 99796 and 99031 in the cross validation subsets and 94605 in the independent test set. The secondary structure contents are similar in all the subsets: about 39% of residues in α-helix, 24% in β-strand and 37% in coil with our assignment and 38%/22%/40% with STRIDE assignment.

3 Hidden Markov Models: application to secondary structure

In a Markovian sequence, the character appearing at position t only depends on the k preceding characters, k being the order of the Markov chain. Hence, a Markov chain is fully defined by the set of probabilities of each character given the past of the sequence in a k-long window: the transition matrix. In the hidden Markov model, the transition matrix can change along the sequence. The choice of the transition matrix is governed by another Markovian process, usually called the *hidden process*. Hidden Markov models are thus particularly useful to represent sequence heterogeneity. These models can be used in predictive approaches: some algorithms like the Viterbi algorithm and the forward-backward procedure allow to recover which transition matrix was used along the observed sequence.

In our case, it is known that different structural classes have different sequence specificity. Intuitively we want to use different Markov chains to model different secondary structures. Figure 2 illustrates the HMM-translation of our secondary structure prediction problem.

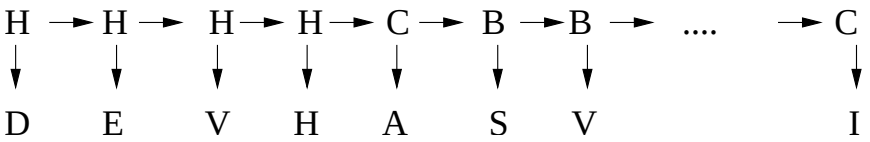

Fig. 2. Secondary structure prediction *via* a hidden Markov model. The upper line represents the secondary structure along a protein sequence: H for a residue in α-helix, B for β-strand, C for coil. The arrows between symbols symbolize the first order dependency of the *hidden process*. The lower line represents the amino-acid sequence of the protein. This is the *observed sequence*. Arrows between the two lines symbolize the dependency between the observed sequence and the hidden chain. The forward/backward algorithm will be used to recover the hidden process from the observed sequence.

The hidden process to be recovered is the secondary structure of the protein. The observed process is the amino-acid sequence. The hidden chain process is a first order Markov chain. Each hidden state is characterized by a distribution of amino-acids. Due to the large alphabet size, the order of the observed chain is 0, which means that amino-acids are independent conditionally on the the hidden process. We use the software called SHOW[1] [Nicolas *et al.*, 2002] to design and train the model and to recover the hidden process. The prediction is achieved with the forward/backward algorithm. Note that this algorithm provides the probability associated to each hidden states at each position.

[1] http://www-mig.jouy.inra.fr/ssb/SHOW/

The simplest model for three-classes prediction is a HMM with three hidden states, each state accounting for a secondary structure class. Parameter estimation of such a model is straightforward because the segmentation is fully determined. But the performance of this model is limited: the Q3 score (proportion of residues with correct prediction) is 58.3%. A random prediction gives a Q3 score equals to 34.5%.

We thus want to design a model that takes into account the specific features of secondary structures.

4 Model of α-helices

A well-characterized sequence motif in α-helices is the amphiphilic motif, i.e., a succession of two polar residues and two apolar residues. This motif occurs when an helix has a side facing the solvent (thus preferentially supporting polar residues) while the other side faces the core of the protein (preferentially supporting apolar residues). This motif is very frequent. With the amino-acids classification; A,V,L,I,F,M,W,C=hydrophobic (h), S,T,Y,N,Q,-H,P,D,E,K,R=polar (p), the motif hhpphh or pphhpp is found in 24% of the helices in our cross-validation set. Glycine (G) residues do not exhibit strong preference for either polar or apolar environment. It is thus considered as a special type of residue and left apart. When reduced to hhpp or pphh, the motif is found in 69% helices. Figure 3 shows the model we propose to take into account the amphiphilic nature of α-helices. States H5 and H6 help to fit the periodicity of an α-helix which is 3.6 residues.

States with hydrophobic preference favour amino-acids A, V, L, I, F, P and M. States with polar preference favour S, T, N, Q, H, D, E, K and R.

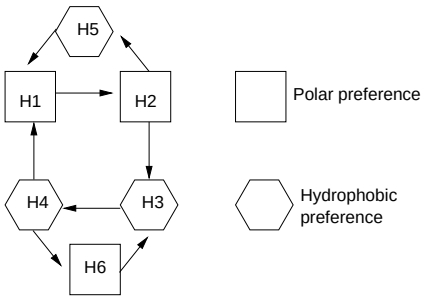

Fig. 3. Model for amphipatic helices

5 Model of β-strands

There is no strong motif characterizing β-strands similar to the amphipatic motif for α-helices. Characteristic motifs are found using a statistical ap-

proach based on exceptional words. A word is over (resp. under)-represented if its frequency in the data is significantly greater (resp. lower) than its expected frequency under some Markovian model. The R'MES software[2] [Bouvier *et al.*, 1999] is dedicated to this task. Amino-acids are grouped as before, the G is put into the hydrophobic group. Sequences of β-strands and α-helices in the cross-validation set are analyzed with R'MES using the Gaussian approximation.

Because the HMM uses a zero order for the observed chain, exceptional words when compared to a zero order Markov model are interesting. Interesting words should also be frequent in absolute (over-represented words are not necessarily frequent) and must not be over-represented in α-helices. We also consider some frequent words, although not over-represented, if they are under-represented in α-helices. Table 1 contains interesting words found in β-strand with R'MES. The over-representation is assessed by R'MES. The relative abundance is evaluated by looking at rank of the word when sorted according to the frequency.

Motif	Occurrence in β-strands	Occurrence in α-helices
hphp	over-represented and frequent	under-represented and not frequent
phph	over-represented and frequent	under-represented and not frequent
pphhh	over-represented and very frequent	under-represented and not frequent
pphph	over-represented and very frequent	under-represented and not frequent
hhhhp	not over-represented, but very frequent	under-represented and not frequent
phhhhp	not over-represented but very frequent	under-represented

Table 1. Interesting motifs in β-strands

Figure 4 shows the model we propose to take into account these words in β-strands. Words hphp and phphp are favoured by the alternation between states b1 and b2. This alternation corresponds to the case of β-strands at the solvent interface with one side facing the solvent and one side facing the core of the protein. The transition from state b4 to itself favours long runs of hydrophobic amino-acids in words pphhh, hhhhp, phhhhp. Long runs of hydrophobic residues are seen when β-strands are buried in the core of proteins. The transition between b2 and b3 favours the apparition of two polar amino-acids surrounded by hydrophobic ones appearing in words pphhh and pphph.

Note that the study of exceptional words on α-helices reveals that the motifs occurring in amphipatic α-helices are over-represented.

[2] http://www-mig.jouy.inra.fr/ssb/rmes/

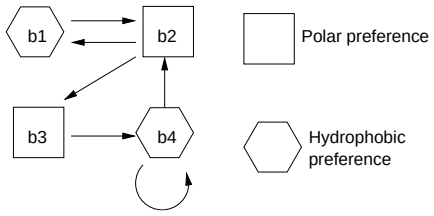

Fig. 4. Model for β-strands

6 Complete HMM for secondary structures

Models of β-strands and α-helices are merged to form a full model of secondary structures.

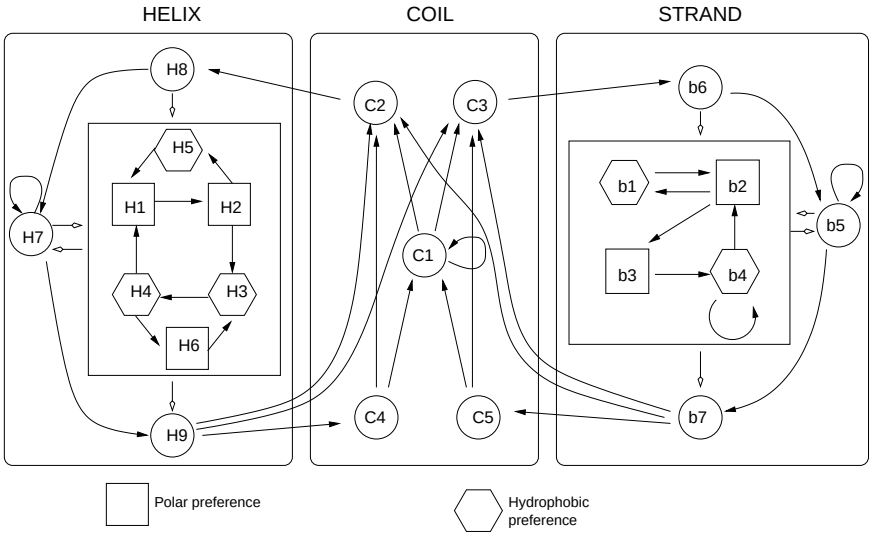

Fig. 5. Full model for secondary structure

Figure 5 shows the full model. Models of α-helices and β-strands integrate informations about frequent/over-represented words, but they don't necessarily reflect the totality of motifs in periodic structures. To allow the presence of β-strands and α-helices that do not fit well in the constrained models, two "generic" states were added (H7, b5). These states show no prior preference for polar or hydrophobic amino-acids. Transitions are allowed between all states of the constrained models and the "generic states". Specific states are added at secondary structure ends (H8, H9, b6, b7), as it is known that there are specific signals such as helix-caps. The coil is not well characterized yet, except the states preceding and following regular secondary structures.

Initial parameters for estimation by the EM algorithm are set as follows:

- Initial transition probabilities are set to $\frac{1}{n}$, with n the number of outgoing states.
- Initial emission probabilities are derived from those obtained on a simple 3-states model. Emission probabilities are manually modified to favour the apparition of polar amino-acids and penalize the emission of hydrophobic amino-acids in polar-preferring states (and vice-versa). No such bias is introduced in other states.

Prediction of the three structural classes (α-helix, β-strand, coil) is achieved by the forward-backward algorithm. The predicted structure is the one with the greatest posterior probability.

7 Integrating information from homologous sequences in the prediction

Protein structures are more conserved than sequences during evolution. Thus different sequences can have the same structure. This information has been successfully used in secondary structure prediction methods [Rost, 2003]. To integrate this information, the prediction is done independently on each sequence of a family. These sequences are detected using a search with PSI-BLAST against a database were the redundancy is reduced to 80% sequence identity. This search generates an average number of 60 sequences per family. Independent predictions are combined with a weighting scheme to generate a prediction for the sequence family using the formula

$$P(state = S/family) = \sum_i \lambda_i \times P(state = S/sequence_i)$$

with $P(state = S/sequence_i)$ provided by the forward-backward procedure and λ_i the weight of sequence i in the family. Sequence weights are computed as proposed in Henikoff and Henikoff [Henikoff and Henikoff, 1994].

Prediction on single sequence provides an accuracy of 65.2% residues correctly classified when compared to our secondary structure assignment, on the cross-validation test set. This score is 65.3% on the learning set and 65.6% on the independent test set. When compared with stride assignment, the accuracy is around 66.3% for all data sets. Hence, we experienced no over-fitting on the training data.

With the family sequence information, the percentage of correct prediction is in the range 71.3 to 72%. Best available methods, that also use sequence families, have achieved accuracy in the range of 78% (reported for reasonnably big datasets on the continuous evaluation server EVA, [Koh *et al.*, 2003]). Thus our results are not fully satisfying yet. However we think that our approach is promising because our model is relatively small,

statistically speaking: the number of independent parameters is only 471. Most of existing methods use neural networks. The number of parameters, when reported, seems to be of the order of thousands [Pollastri *et al.*, 2002]. Moreover, the graphical nature of hidden Markov models allows intuitive data modeling. Along this line, an important perspective of this work is to introduce a geometrical description of coil. The coil class represents about 50% of residues in proteins. Even a perfect three state prediction would leave half of the data with no structural clue. We also think that the sequence family information could be taken into account more efficiently that it is done here. This is another of our perspectives.

References

[Bouvier *et al.*, 1999]A. Bouvier, F. Gélis, and S. Schbath. *RMES : Programs to Find Words with Unexpected Frequencies in DNA Sequences, User Guide (in french)*, 1999.

[Brenner *et al.*, 2000]S.E. Brenner, P. Koehl, and M. Levitt. The astral compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–6, Jan 2000.

[Frishman and Argos, 1995]D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–79, Dec 1995.

[Henikoff and Henikoff, 1994]S. Henikoff and JG. Henikoff. Position-based sequence weights. *J Mol Biol*, 243(4):574–8, Nov 1994.

[Koh *et al.*, 2003]I.Y. Koh, V.A. Eyrich, M.A. Marti-Renom, D. Przybylski, M.S. Madhusudhan, N. Eswar, O. Grana, F. Pazos, A. Valencia, A. Sali, and B. Rost. Eva: evaluation of protein structure prediction servers. *Nucleic Acids Res*, 31(13):3311–5, Jul 2003.

[Nicolas *et al.*, 2002]P. Nicolas, A.S. Tocquet, and F. Muri-Majoube. *SHOW : Structured HOmogeneities Watcher. User Guide*, 2002.

[Pollastri *et al.*, 2002]G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 2002.

[Rost, 2003]B. Rost. Prediction in 1d: secondary structure, membrane helices, and accessibility. *Methods Biochem Anal*, 44:559–87, 2003.