Markov Models for Classification of Protein Helices

Nancy Ruonan Zhang
Biochemistry 218
June, 01

Introduction

The amphipathic alpha helix is a common secondary structural motif in globular proteins. An amphipathic alpha helix is defined as an alpha helix with opposing hydrophobic and hydrophillic faces oriented along the long axis of the helix. Through the analysis of hemoglobin and myoglobin, Perutz et. al. (1990) showed that amphipathicity can be detected through sequence signals: Non-polar amino acids appear approximately every 3.6 residues in the linear sequence, making one face of the folded helix hydrophobic. Experimental studies have also shown that amphipathicity is a major driving force in the formation of helices and the folding of protein structures (Scheraga, 1985; Lyu et.al., 1990). The objective of this study is to characterize the periodicity in amphipathic helices using Markov models, and to assess the effectiveness of these models in discriminating between different helices of different types.

The amphipathicity of an alpha helix can be determined by fitting its residues on a "Schiffer-Edmundson" helical wheel diagram (Schiffer & Edmundson, 1967). The helical wheel diagram provides a top view of the helix, with the residue side-chains projected onto a plane perpendicular to its long axis. A similar method (Lim, 1978) projects the side-chain positions on a "helical net" which imitates a cylindrical surface wrapped around the long axis of the helix. Both of these methods are easy to use, but suffer certain shortcomings. It has been shown (Flinta et. al., 1983) that the helical wheel representation has a tendency to over-estimate helical amphipathicity. Furthermore, these methods are inadequate for studying short helices (< 6 residues) and are tedious for the analysis of long sequences.

Fourier analysis offers a more quantitative approach to measuring the amphipathicity of alpha helices. The idea is to view the hydrophobicity (or solvent accessibility, if data is available) of the residues in the protein sequence as a discrete linear signal, and analyze it using the Fourier transform in the frequency domain. The plot of these frequencies between 0 and 180 degrees has been traditionally called "Eisenberg Plots." Eisenberg et. al. (1984) was one of the first to use this method to identify significant Fourier intensities at around 100 degrees, which corresponds to the 3.6 residue per turn of alpha helices. Eisenberg et. al. (1982) also introduced the mean helical hydrophobic moment, which gave a quantitative interpretation of the helical wheel. In frequency domain, Eisenberg's definition of the mean hydrophobic moment simply translates to the modulus of the discrete Fourier transform at 100 degrees.

More useful to this study are the short-range signals that can be detected in amphipathic helices, since only such signals can be detected using Markov models of low order. An accumulation of hydrophobic triplets at positions n, n+3, n+4 and at positions n, n+1, n+4 were found by Palau and Puigdomenech (1974) to have a stabilizing effect on amphipathic helices. More recently, a study by Negrete et. al. (1998) confirmed these findings and further concluded that these triplet signals are a universal feature found in amphipathic helices of globular proteins, whatever the overall architecture may be.

The periodic pattern in amphipathic helices have proved useful in secondary structure prediction and helix classification. Wako and Blundell (1994) combined amino acid substitution patterns, helix capping signals, and Fourier transform approaches in a secondary-structure prediction algorithm that reached 77% accuracy. In a later study, Zhu & Blundell (1996) made more explicit use of the size of the solvent-inaccessible face of an alpha helix to predict not only the position of a secondary structure in a protein sequence but also its orientation with respect to the core of the protein. In effect, the algorithm of Zhu & Blundell not only predicts the presence of helices, but also classifies them according to their orientation within the protein. A more functional classification of helices was given by Segrest et. al. (1990), which used the mean hydrophobic moment and other techniques to group amphipathic helixes in to seven classes of different physical-chemical and structural properties.

One critical assumption that many studies make about alpha helices is that they have the ideal structure defined by the classical work of Pauling et. al., i.e. that the alpha helix is a regular helix with 3.6 residues per turn. The construction of the Schiffer-Edmundson wheel and Lim's helical net, especially, rely heavily upon this assumption of regularity. In reality, however, up to 80% of helices in globular proteins deviate from this ideal (Barlow & Thornton, 1988). Many long helices of over four turns are kinked, caused by the presence of a proline in the interior of the helix. A majority of helices also exhibit a slight degree of curvature of the helical axis (~60

Angstrom radius), which, in amphipathic helices, generally points towards the hydrophobic core (Blundell et. al., 1983). Cornette et. al. (1987) found that the dominant frequency for alpha helices is at 97.5 degrees rather than 100 degrees, suggesting that the helix is slightly more open than previously thought, with the number of residues per turn closer to 3.7 than 3.6. Furthermore, it is commonly known that hydrophillic residues do occur in the buried face of helices, and that hydrophobic residues are sometimes exposed to the solvent. These deviations from the ideal pose a problem to any model-dependent study that rely heavily on the helical wheel or helical net representation of the alpha helix.

This study takes an unsupervised machine learning approach to analyzing the Markov dependency between residues in alpha helices. An ideal helix assumption is used to aid the design of model structure and to analyze the results, but is not directly incorporated in to the model or used in parameter estimation. The unsupervised, exploratory parameter estimation approach allows data-dependent learning of parameters that capture not only the overall trend in different classes of helices, but also any deviation from the trend. Several different stochastic models of alpha helices are explored, each of which assumes a different pattern of dependency between nearby residues within the linear sequence. A clustering method based on the expectation-maximization (EM) algorithm is used to approximate the maximum-likelihood parameters to these models, as well as to find the optimal classification of the data. The models are evaluated based on their maximum likelihood scores and their ability to classify helices into groups of different levels of amphipathicity.

By exploring several different Markov models for the alpha helix, this study is able to assess the importance of different inter-residue dependencies to the classification and representation of helices. What type of model is best for discriminating between amphipathic and non-amphipathic helices? In contrast, what type of model is best for representing alpha helices of different degrees of amphipathicity? We'll see in the results section that the model that performs the best for the former task may not perform as well for the latter task. What is the simplest model that is can capture the periodic signal in alpha helices and store that information in its parameters? We'll see that, with a dependence window of only 3 residues, a model is capable of capturing the periodicity of 3.6 residues in amphipathic alpha helices. Using Monte-Carlo simulations, we will attempt to answer the question, "How much information is really stored in the model parameters?" Finally, we'll define a distance measure on model space and use it to characterize the similarity between any two parameterizations of the same model. This allows us to examine the nature of the likelihood function on the parameter space and to evaluate the confidence of our EM parameter estimates.

Methods and Model Descriptions

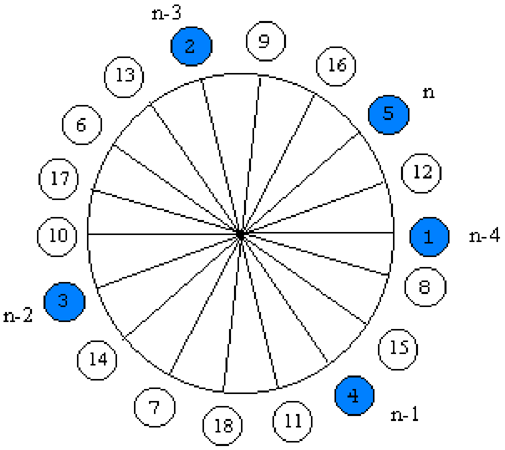


Figure 1

Figure 1 shows the relative locations of five linearly sequential residues on a helical wheel projection. Assuming an amphipathic helix, we would expect that the hydropathy value of residue n would correlate negatively with that of residue n-2, since they are on opposite sides of the helix. We would also expect a positive correlation between the hydropathy values of residues n and n-4, due to their proximity in the helix wheel projection. The correlation between the hydropathy of residues n-3 and n-1 and that of residue n depends on the size of the inaccessible face of the helix, and the position of residue n relative to that face. These are the dependency information that we are trying to capture in this study. In all of the Markov models analyzed, a certain dependency is assumed between the nth residue and the n-1, n-2, n-3, and n-4-th residues, with n running along the entire length of the interior of the helix. As shown in figure 2, the interior of the helix

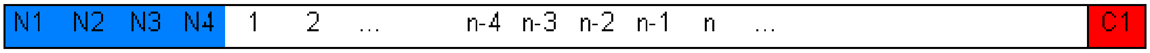


Figure 2

is defined to start 4 residues from the n-terminal end and 1 residue form the c-terminal end of the protein sequence. The rationale for this partitioning can be found in the analysis of Schmidler et. al. (2000), which shows that amino acid composition at positions 1, 2, 3, and 4 residues from the n-terminal and 1 residue from the c-terminal differs significantly from the overall amino acid composition.

To reduce the parameter space of the models, the Markov dependency at the helix interior is specified in terms of hydropathy class instead of in terms of amino acid value. Three hydropathy classes are defined: hydrophobic, hydrophillic, and neutral. The amino acids are grouped into these three classes based on the Kyte-Doolittle hydropathy scale as shown in table 1. If we assume that the distribution is multinomial and that each residue is probabilistically dependent only on the previous two residues, then there would be $3^2 * 2 = 18$ parameters in the model, instead of $20^2 * 19 = 7600$ parameters as in the case where the inter-residue dependence is specified in terms of amino acid value. Given the hydropathy class, the models assume that the amino acid value follows a stationary multinomial distribution, the parameters for which must also be estimated by the learning process. The n- and c- terminal residues are assumed to be independently distributed with their own distinct multinomial distributions.

AA Code	Kyte-Doolittle hydropathy	Hydropathy class
Arg	-4.5	Hydrophillic
Lys	-3.9	Hydrophillic
Asn	-3.5	Hydrophillic
Asp	-3.5	Hydrophillic
Gln	-3.5	Hydrophillic
Glu	-3.5	Hydrophillic
His	-3.2	Hydrophillic
Pro	-1.6	Neutral
Tyr	-1.3	Neutral
Trp	-0.9	Neutral
Ser	-0.8	Neutral
Thr	-0.7	Neutral
Gly	-0.4	Neutral
Ala	1.8	Hydrophobic
Met	1.9	Hydrophobic
Cys	2.5	Hydrophobic
Phe	2.8	Hydrophobic
Leu	3.8	Hydrophobic
Val	4.2	Hydrophobic
Ile	4.5	Hydrophobic

Table 1

Five different models are studied, of varying degrees of complexity. The simplest model (IID), which serves as control, assumes that each residue in the interior segment of the helix is independent and identically distributed. The model of the next level of complexity (D(n-1))assumes that the interior residues are first-order Markov, (i.e. each n-th residue is dependent only on the n-1-th residue, from which it is separated by an 100-degrees arc in the helical wheel projection). Two models are studied that assume second-order markovicity for the internal residues: one assumes dependence of each n-th residue on the n-1 and n-2-th residues (D(n-1, n-2)), and the other assumes dependence on the n-1 and n-4-th residues (D(n-1, n-4)). The critical difference between these two models is that the residue n-2 is on the opposite side of residue n in the helical wheel projection, while residue n-4 is on the same side. Finally, the model of the highest level of complexity (D(n-1, n-2, n-3)) assumes that the interior residues are third-order Markov. Table 2 summarizes the dependency assumptions of the models.

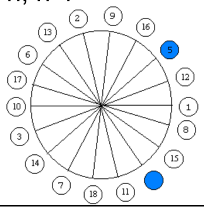
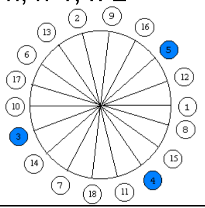
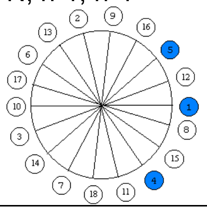
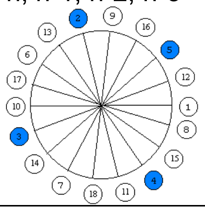
Model Acronym	IID	D(n-1)	D(n-1, n-2)	D(n-1,n-4)	D(n-1, n-2,n-3)
Dependence Relationship Assumed as Projected on to the Helical Wheel	None	n, n-1  2 9 13 16 5 6 12 17 1 10 8 3 15 14 7 18 11	n, n-1, n-2  2 9 13 16 5 6 12 17 1 10 8 3 15 14 4 7 18 11	N, n-1, n-4  2 9 13 16 5 6 12 17 1 10 8 3 15 14 4 7 18 11	n, n-1, n-2, n-3  2 9 13 16 5 6 12 17 1 10 8 3 15 14 4 7 18 11

Table 2

Model parameters are estimated using the EM algorithm for mixture distributions (see Cheeseman et. al. 1988a & 1988b or Duda & Hart for more in-depth description of the algorithm). The prior probability distributions for all of the parameters are assumed to be Dirichlet. The number of clusters *c* is considered a fixed, known value. The models used for all clusters abide by the same Markov assumption, but through the learning process gain different parameters. The true classifications for the protein sequences are not known, as according to the unsupervised learning paradigm. The pseudo code of the EM algorithm is as follows (*D* = complete data set of helix sequences, *M* = mixture model):

```
1. Randomly initialize model parameters according to the prior distribution.
2. Repeat until gain in $P(D|M)$ between iterations is smaller than some predefined
   threshold:

      (i)   Classify helices according to the current model parameters. That is, for
            each helix $h$ and cluster $i$, calculate $P(i|h)$.
      (ii)  Use the resulting classification to re-estimate the model parameters for
            each cluster that would maximize the conditional likelihood of the
            complete data set given the mixture model (i.e. $P(D|M)$).

3. Report final model.
```

To alleviate the problem of local maxima, random restart is employed to explore more of the parameter space. Hence, every time a final model is reported, steps 1, 2, and 3 are repeated with a different random initialization of the model parameters. For this study, random restart is employed 30 times before a best model is selected.

In order to better understand the parameter space, a distance measure must be defined to describe the degree of similarity between any two model parameterizations. If, after every random restart, the EM algorithm reports a model that is a negligible distance away from the previously found models, then it is highly possible that most of the parameter space has already been explored, and that the globally optimal model is among the current set of reported models. The distance measure also makes possible the evaluation of the complexity of the parameter space by allowing a quantitative comparison between models of different likelihood scores. The distance measure for IID models is defined as follows:

Let $p_1(x, k) = P(\textit{amino acid } x, \textit{ class } k)$ according to model 1
$p_2(x, k) = P(\textit{amino acid } x, \textit{ class } k)$ according to model 2

$x \in$ {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}
$k \in \{1\ldots c\}$

Then $d(\textit{model 1, model 2}) = max_q D(p_1(x, k)||q(x, k))$,
where q is any permutation of $p_2(x, k)$ *with respect to k,*
and $D(p||q)$ *is the Kullback Leibler Divergence*

The distance measure for the other models are defined similarly. The idea is to find the pair-wise match between the classes of the two models so as to minimize the KL-divergence between the corresponding distributions.

Three methods are used to assess the amphipathicity of a single or a set of helices: The "Eisenberg plot", the hydrophobic moment, and the size of the hydrophobic face. As described in the introduction, The "Eisenberg plot" is a plot of the hydrophobicity of the protein sequence in frequency domain. Given the hydrophobicity values of each residue in the helix $\{h_i\}_{i = 1\ldots n}$, the "Eisenberg plot" plots $\{f_\omega\}_{\omega=0\ldots180}$, the Fourier transform of $\{h_i\}_{i = 1\ldots n}$. The frequency vector $\{f_\omega\}$ is calculated as follows:

$$f_\omega = \{ [\textstyle\sum_{k=1\ldots n} h_k \cos(k\omega)]^2 +[\textstyle\sum_{k=1\ldots n} h_k \sin(k\omega)]^2 \}^{1/2} ,$$

$$\omega = 0\ldots180$$

A peak in the "Eisenberg plot" around 100 degrees had often been used as a good indicator of an amphipathic helix. Note that although the Eisenberg plot shows the strength of any periodic signals within a helix, it does not give much information about size of the hydrophobic face.

To calculate the hydrophobic moment and the size of the hydrophobic face of a helix, its residues must first be mapped from the linear sequence representation into the helical wheel representation. For an alpha helix with less than 18 residues, linear interpolation is used to fill in the missing values in the wheel. For an alpha helix with more than 19 residues, the average is used at position(s) where there is more than one residue. The hydrophobic moment is calculated via the method described in Segest et. al. (Segrest, Jones & Anantharamaiah, 1992): For each residue projected on to the wheel, multiply the unit vector that extends radially from the wheel's center to the position of the residue on the wheel's circumference by the hydropathy value of the residue. The magnitude of the sum of the vectors is the helical hydrophobic moment. The size of the hydrophobic face is measured by the length of the longest stretch of non-hydrophillic residues along the wheel. For example, if the helix is non-amphipathic and composed completely of hydrophobic residues, then the hydrophobic arc-length would be almost 18. On the other hand, if the helix is completely exposed, it is most likely that the hydrophobic arc length would be close to 0. Amphipathic helices have arc-length between these two extreme values.

A simple Monte-Carlo type approach is used to access the statistical significance of any feature of the models: For each sequence in the database, randomly permute its residues while conserving their relative frequencies. Then, use the same EM algorithm to train models using this scrambled database. Any periodicity or between-class divergence that is encoded by the parameters of models trained in this way must have been due to chance. Many such randomized models were generated and used to estimate the mean and standard deviation of the null-distribution for the feature. A critical assumption is made that the null-distribution is normal, and the z-test is used to test the significance of the feature.

Results & Discussion

I. *Overall assessment of model performances.*
Figure 3 plots the Akaike Information Criterion (AIC) (Akaike, 1974) of the different model types for increasing number of classes. The AIC for each model is calculated as follows:

AIC = 2* [(number of parameters) – (model likelihood)]

While the conditional likelihood score increases indefinitely with the increase of the number of classes and the order of markovicity, the AIC criterion penalizes models with large parameter spaces. Thus, the AIC offers a quantitative way of comparing the performance of models with different numbers of parameters. As expected, Figure 3 shows that IID models perform the worst, followed by D(n-1). The AIC curves for D(n-1, n-2, n-3), D(n-1, n-2) and D(n-1, n-4) are relatively the same. This indicates that the extra n-3-rd residue considered by the D(n-1, n-2, n-3) model does not improve its performance, by the standard of the AIC. Also noticeable is that the higher order models are paraboloid in shape. This is because these models have a larger number of parameters per class, and thus their model complexity increases much faster than that of the IID or D(n-1) model. The AIC curves for all of D(n-1, n-2), D(n-1, n-4), and D(n-1, n-2, n-3) reach a minimum at 4 classes, suggesting that any model with more than four classes may be over-parameterized.

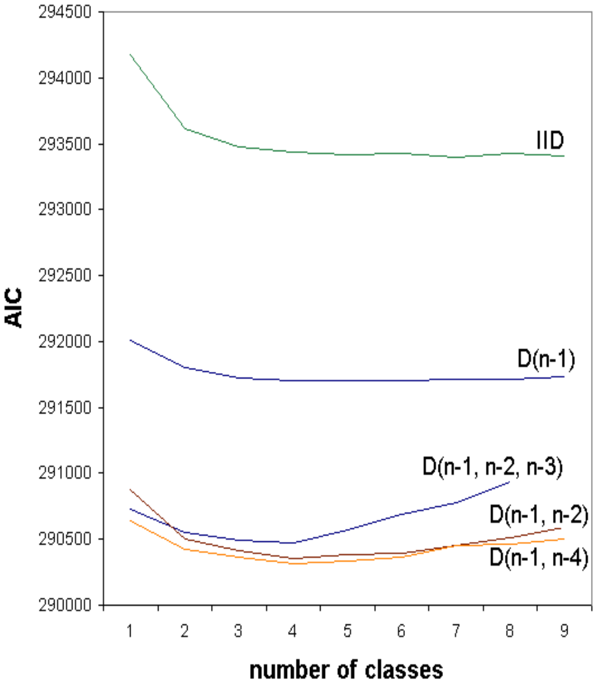


Figure 3

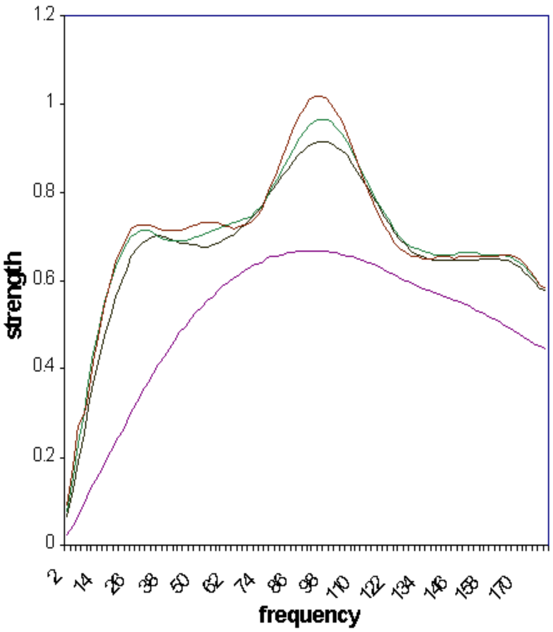


Figure 4

I. *Amphipathic signals in helical sequences can be captured by any Markov model of order greater than 2.* Although the alpha helix has 3.6 periodicity and helix-stabilizing interactions have been detected between residues as far as 4 apart, the results of this study indicate that even a second-order Markov model can detect amphipathic signals in helices. Figure 4 shows the mean Eisenberg plot for a D(n-1, n-2) 4-class model. The between-class difference in the mean modulus at 100 degrees has been tested to be statistically significant (P-value > 0.99)

The plots in figure 4 were calculated using real protein sequences from the original database. Hence, we can not yet infer that the strong signal at 100 degrees is in fact encoded in the model itself. How much information is actually captured and stored in the model parameters? Five-thousand "artificial" amino acid sequences are simulated Monte-Carlo-style using each of the class-conditional-distributions of the model being analyzed. Any statistically significant periodic signals found in this simulated set must have been encoded by the model parameters, and thus "learned" by the model through the EM algorithm. Figure 5 shows the results of these simulations for D(n-1), D(n-2), and D(n-1, n-2, n-3).

The Eisenberg Plots for all of the class-distributions within the D(n-1) model are flat and show no dominant frequency. For the D(n-1, n-2) model, the classes have noticeably different plots, with that of classes 1 and 3 having statistically significant peaks at approximately 97 degrees. The plots suggest that class 1 contains the most strongly amphipathic helices, class 2 contains the non-amphipathic helices, and class 3 is somewhere in between the two. These results are statistically significant (P-value > 0.99) The results for the D(n-1, n-2, n-3) model are even better, as the difference in peakedness between the class plots is even more noticeable. This has been expected, since as the order of the model increases, more information can be stored in its parameters, and thus we would expect it to perform better.

It may at first seem counter-intuitive that a Markov model of only second order can capture signals of periodicity greater than 3. The fact that this is possible can be shown through a trivial exercise. Consider the deterministic second order Markov chain with the following transition rules:

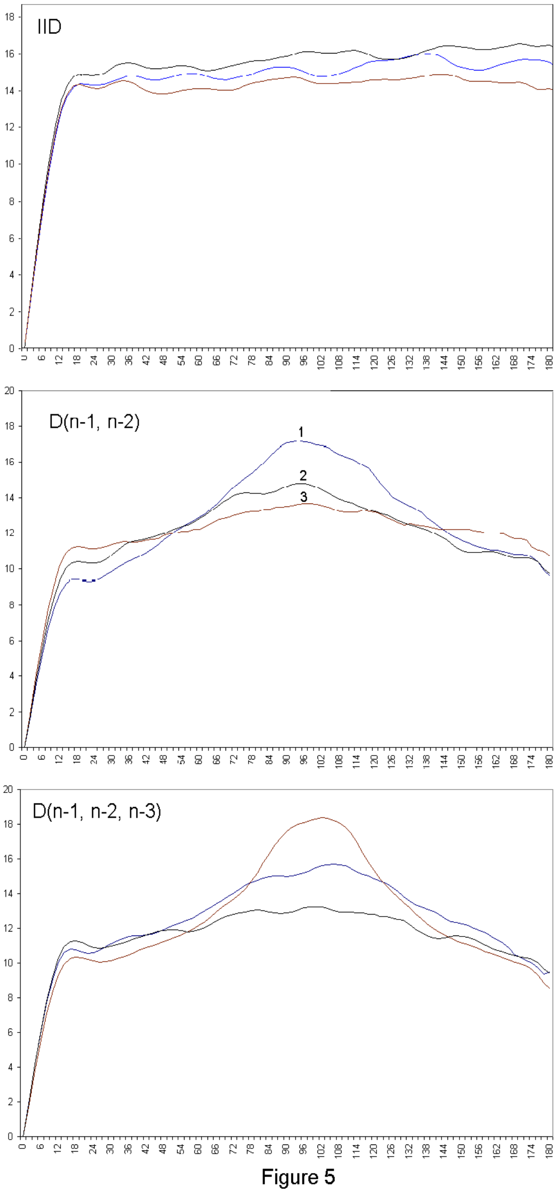


Figure 5

n-2	n-1	n
hydrophobic	hydrophobic	hydrophillic
hydrophobic	hydrophillic	hydrophillic
hydrophillic	hydrophobic	hydrophobic
hydrophillic	hydrophillic	hydrophobic

If the two residues at the start of the sequence are both hydrophobic, then the simulated sequence would have the following pattern:

hydrophobic
hydrophobic
hydrophillic
hydrophillic
hydrophobic
hydrophobic
hydrophillic
hydrophillic
hydrophobic
hydrophobic
hydrophillic
hydrophillic
…..

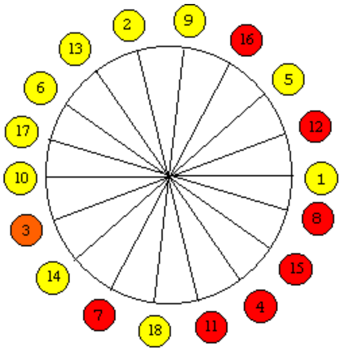


Figure 6

Figure 6 shows the projection of this simulated sequence on to the helical wheel. For this simple, deterministic case, the simulated sequence has a periodicity of 4 and the projected helix wheel diagram clearly shows an amphipathic helix. The D(n-1, n-2) model parameters for amphipathic helices are simply a probabilistic version of the above trivial case, with added noise and three hydropathy classes instead of two. By the position of the residues n, n-1, and n-2 on the helical wheel, it is logical that if the hydropathy class of residue n correlates negatively with that of residue n-2, then the resulting sequence would be amphipathic.

II. *The D(n-1, n-2) model is more effective for discriminating between different amphipathic helices, while the D(n-1, n-4) model is more effective for discriminating between amphipathic and non-amphipathic helices.* As noted in the Methods section, the critical difference between D(n-1, n-2) and D(n-1, n-4) is in the location of the n-2-nd and the n-4-th residues on the helical wheel, relative to that of the n-th residue. While residue n-2 is on the opposite face of residue n, residue n-4 is on the same face. Furthermore, the arc spanned by residues n, n-1, and n-2 is 11 residues long, while that spanned by residues n, n-1, n-4 is only 6 residues long. Thus, intuitively, D(n-1, n-4) should not be able to differentiate between an amphipathic helix with inaccessible face of size 6 residues from that with inaccessible face of size more than 6 residues, because its projected helix wheel “reading frame” is only 6 residues long. However, although the projected “reading frame” of D(n-1, n-4) is smaller than that of D(n-1, n-2), it is more detailed, since 3 residues within this size-6 frame (50% of the residues) is observed. By contrast, the projected reading frame of D(n-1, n-2) is larger (11 residues) but less detailed (3/11 = 27% residues is observed). The effect of this difference on classification results is shown in table 3.

Model Type	D(n-1, n-2)	D(n-1,n-4)	D(n-1, n-2, n-3)
Mean Arc Length	Class 1: 6.10	Class 1: 6.46	Class 1: 5.91
	Class 2: 6.93	Class 2: 6.70	Class 2: 7.23
	Class 3: 7.83	Class 3: 7.19	Class 3: 8.14
Mean Hydrophobic Moment	Class 1: 12.95	Class 1: 11.14	Class 1: 10.92
	Class 2: 13.39	Class 2: 11.25	Class 2: 11.31
	Class 3: 14.21	Class 3: 14.46	Class 3: 14.22
	Class 4: 14.51	Class 4: 15.00	Class 4: 15.18

Table 3

Three models are analyzed (D(n-1, n-2, n-3), D(n-1, n-2), D(n-1, n-4)), each assuming the presence of 4 statistically separable classes of helices. The results for D(n-1, n-2, n-3) serve as control, since the Markov dependency assumed by this model, when projected on to the helical wheel, is the sum of that assumed by D(n-1, n-2) and D(n-1, n-4). Each model found three statistically significant (P-value > 0.98) amphipathic helix classes and one non-amphipathic helical class. The between-class difference in mean hydrophobic-face size for models D(n-1, n-2, n-3) and D(n-1, n-2) are statistically significant, but that for model D(n-1, n-4) is not. This result suggests that the triplet n, n-1, n-2 is more informative for distinguishing between helices of different degrees of amphipathicity. However, comparison of the hydrophobic moment produced by the two models reveals that D(n-1, n-4) is more effective at capturing the periodicity information in helices. The hydrophobic moment produced by the class-conditional-distributions of D(n-1, n-4) cover a wider range and are better separated from each other. The most amphipathic class found by D(n-1, n-2) has hydrophobic moment of 14.51, while that found by D(n-1, n-2) has hydrophobic moment 15.00. These results have borderline statistical significance (P-value > 0.93). Thus, further testing is needed to confirm their validity.

IV. The likelihood functions for the non-zero order Markov models have many local maxima. As described in the methods section, 30 iterations of the EM algorithm were performed, each from a different random initialization of parameters. Each run generates a “best” parameterization of the given model, among which the parameterization that gives the overall highest likelihood score *M** is selected. The distance between the other reported models and M* is calculated and plotted against the decrease in their likelihood scores. Figure 7 shows the plots for the IID and D(n-1, n-2) models.

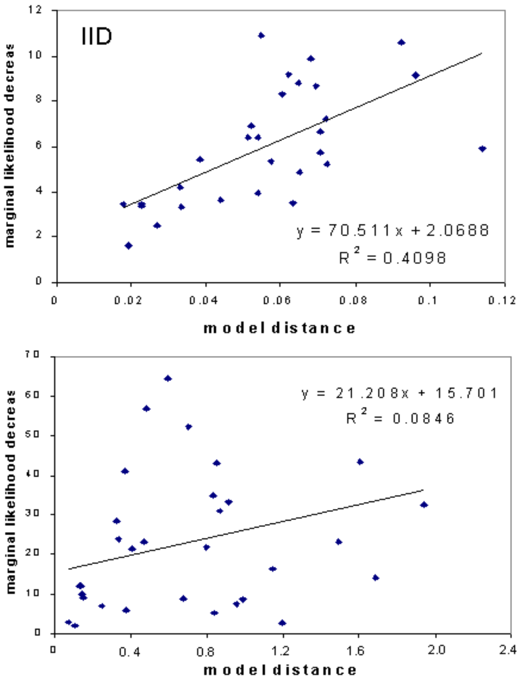


Figure 7

The data suggests that the likelihood scores for the IID models are inversely related to their distance from the maximum likelihood score. All of the 30 IID models are within a short distance away from the maximum likelihood model. This suggests that the IID likelihood function is unimodal, and that the best model that we have found so far is indeed the global optimum. However, the data for D(n-1, n-2) shows no significant relationship between model score and model distance. The reported models are also much further apart than in the IID case. Thus, comparatively, the likelihood function for D(n-1, n-2) is much more complex and most likely contains many local maxima over the parameter space. Based on this fact, 30 iterations of random restart may not nearly be enough to guarantee that the entire parameter space has been explored. This is expected, since usually the number of iterations needed to explore the entire parameter space scales exponentially with the number of parameters.

Conclusion

This study presents a method that uses Markov models to classify alpha helices in to amphipathic and non-amphipathic classes. The second- and third- order Markov models found using the EM algorithm conferred interesting insights on the inter-residue dependence in amphipathic alpha helices, despite the fact that they may not be globally optimized. Considering that the models were trained using unlabeled data, the fact that they were able to find classes with significant differences in amphipathicity is evidence that low-order Markov amphipathic signals in helix sequences are quite strong.

The helix classification models found in this study can be incorporated into the secondary structure prediction algorithm designed by Schmidler. et. al. They can also be used to detect the amphipathic helix motif in protein sequences.

References

Akaike, H. 1974 *IEEE Trans. Autom. Contr.*, 19:716-723.

Barlow D. J., Thornton J. M. 1988. *J. Mol. Biol.*, 201:601-619.

Blundell T.L., Barlow D., Borkakoti N., Thornton J. 1983. *Nature*, 306: 281.

Blundell T.L., Zhu Z.-Y. 1995. *Biophys. Chem.*, 55:167-184.

Cheeseman P., Kelly J., Self M., Stutz J., Taylor W, Freeman D. 1988a. Autoclass: A Bayesian Classification System. In *Proceedings of the Fifth International Conference on Machine Learning*.

Cheeseman P., Self M., Kelly J., Stutz J. Taylor W, Freeman D. 1988b. *Seventh National Conference on Artificial Intelligence*, 607-611.

Cornette J.L., Cease K.B., Margalit H., Spouge J.L., Berzofsky J.A., DeLisi C. 1987. *J. Mol. Biol*. 195, 659-685.

Eisenberg D., Weiss R.M., Terwilliger T.C. 1982. *Nature* 299, 371-374.

Eisenberg D., Weiss R.M., Terwilliger T.C. 1984. *Proc. Natl. Acad. Sci.,* 81: 140-144.

Flinta C., Von Heijne G., Johansson J., 1983. *J. Mol. Biol.,* 168: 193-196.

Jones M.K., Anantharamaiah G.M., Segrest J.P. 1992. *J. Lipid Res.*, 33:287-296.

Kyte J., Doolittle R.F. 1982. *J. Mol. Biol.,* 157: 105-132.

Lim V.I. 1978. *FEBS Lett*., 89: 10-14.

Lyu P.C., Liff M.I., Marky L.A., Kallenback N.R. 1990. *Science*, 250: 669.

Negrete, J.A., Vinuales, Y., Palau, J. 1998. *Protein Sci.*, 7: 1368-1379.

Palau J., Puigdomenech P. 1974. *Int J. Pept Protein Res*, 19:394-401.

Perutz M.F. , Kendrew J.C., Watson H.C. 1965. *J. Mol. Biol.*, 13:669.

Segrest J.P., Loof H.D., Dohlman J.G., Brouillette C.G., Anantharamaiah G.M. 1990. *Proteins*, 8: 103-117.

Scheraga H.A.1985. *Proc. Natl. Acad. Sci., USA*, 82 5585.

Schiffer M., Edmundson A.B. 1967. *Biophys. J.*, 7:121.

Schmidler S.C., Liu J.S., Brutlag D.L.. 2000, *Bayesian Segmentation of Protein Secondary Structure* (Work in Progress)

Zhu, Z.-Y., Blundell, T.L. 1996. *J. Mol. Biol*., 260: 261-276.

Wako H., Blundell T.L. 1994. *J. Mol. Biol*., 238: 682-692.