Multicoil-HMM: improved prediction of coiled-coil oligomer state from sequence

by

Jason Trigg

S.B. Electrical Engineering and Computer Science, M.I.T., 2010

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

August, 2010
[September 2010]

Author___ _________________________
Department of Electrical Engineering and Computer Science
August 20, 2010

Certified by__ ___________________
Professor Bonnie Berger
Thesis Supervisor

Accepted by________ ______________________
Dr. Christopher J. Terman
Chairman, Department Committee on Graduate Theses

Multicoil-HMM: improved prediction of coiled-coil oligomer state from sequence
by
Jason Trigg

Submitted to the
Department of Electrical Engineering and Computer Science

August 20, 2010

In Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

The Multicoil-HMM algorithm offers improved prediction of coiled-coil oligomerization state. The algorithm combines the pairwise correlations of the Multicoil method with the flexibility of HMM methods. The resulting method incorporates predictors deemed important by a multinomial logistic regression to distinguish between the dimer, trimer and non-coiled coil oligomerization states. The Multicoil-HMM algorithm shows significantly improved oligomer state prediction over a retrained Multicoil algorithm, which is currently the state-of-the-art. The general strategy of using multinomial regression on predictors that can be simulated by HMMs while abandoning the probabilistic interpretation of HMMs may be useful in other machine learning applications.

Acknowledgements

I would like to thank my thesis advisor Bonnie Berger. I am grateful to Amy Keating for her advice and extensive knowledge of coiled coils. Thanks to Karl Gutwin for the construction of the coiled coil and negatives databases, as well as the description of those databases. Also thanks to Karl Gutwin and Andrew McDonnell for retraining the original Multicoil method on this new data.

Introduction

As the rate of protein sequencing continues to outpace structure identification, there is a need for algorithms to predict protein structure from sequence. In the case of coiled coils, the process is eased somewhat by the period seven heptad structure, along with the fact that interactions in the coiled coil structure are mainly localized, unlike, for example, beta sheets. Still, coiled coil prediction is problematic, particularly distinguishing between different oligomerization states. We examine current methods, and finding them deficient, make a more general predictor that combines their strengths. The Multicoil-HMM method first removes the constraint that HMM emission and transition probabilities be normalized. Then the unnormalized emission probabilities are carefully chosen such that a path through states corresponding to a coiled coil has a probability related to a linear combination of basic coiled coil predictors. Lastly, the optimal linear combination of basic predictors is generated by training a multinomial logistic regression.

Background

Coiled Coils

The coiled coil structure is formed when several α-helices wrap around each other with a left-handed superhelical twist. In an α-helix, the amino acids wrap at a rate of 3.6 residues per rotation. However, in the coiled coil conformation, the α-helices are wrapped slightly tighter, at a rate of 3.5 residues per rotation. That is, after 7 residues, the alpha-helices making up the coiled coil have made exactly 2 rotations. Thus there is a heptad structure to coiled coils, denoted $(abcdefg)_n$. As shown in figure 1, the a and d positions from the two α-helices interact, and thus we find that the residues in these positions are hydrophobic while residues in the other positions are hydrophilic.

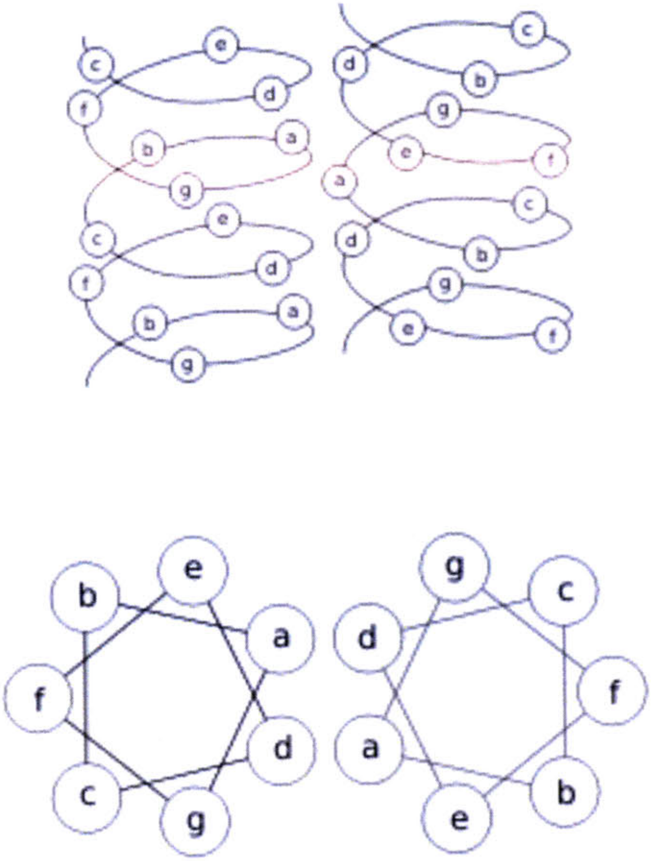

Figure 1: Diagram of a dimeric coiled coil with the heptad positions labelled. Every seven residues, the alpha helices make exactly two rotations. All the residues in the a and d positions interact with their counterparts on the opposite helix, and thus are generally hydrophobic.

Prior methods

The heptad pattern of coiled coils has allowed for successful sequence-based prediction of the coiled coil motif. Prediction of coiled coil structure has been mainly conducted through the position specific scoring matrix (PSSM) and hidden Markov model (HMM) methods.

PSSM methods

The first methods for prediction of coiled coils looked, residue by residue, for some region around that residue with high coiled coil propensity. For each residue, the PSSM method considers all windows of a fixed width n (usually 21 or 28 residues) containing that residue, and every heptad assignment for the chosen residue. The PSSM method gives the entire window a score, tries all n*7 window alignments/heptad choices for the given residue, and assigns the residue the maximum score of any of the windows containing that residue. For example, in figure 2, we are currently considering the indicated glutamine. The example uses a length 21 PSSM window, and several possible window locations and heptad alignments are shown. The PSSM method scores the coiled coil propensity of all these alignments and assigns the residue the maximum of those propensities.

PSSM methods have their drawbacks. First of all, the fixed window size is suboptimal - for coiled coils shorter than the length of the window, the rest of the window is filled with non-coil amino acids; this noise attenuates the signal. On the other hand, many coiled coils are longer than the window, and the PSSM methods cannot make use of any evidence farther away than one window length when predicting whether a residue is in a coiled coil.

Choosing the best scoring method for the windows is non-trivial. The prediction function includes a step in which we maximize over possible windows, so the prediction is not a differentiable function of the parameters of the scoring function. This complicated objective function makes it hard to optimize over scoring methods. As a result of the difficulty in

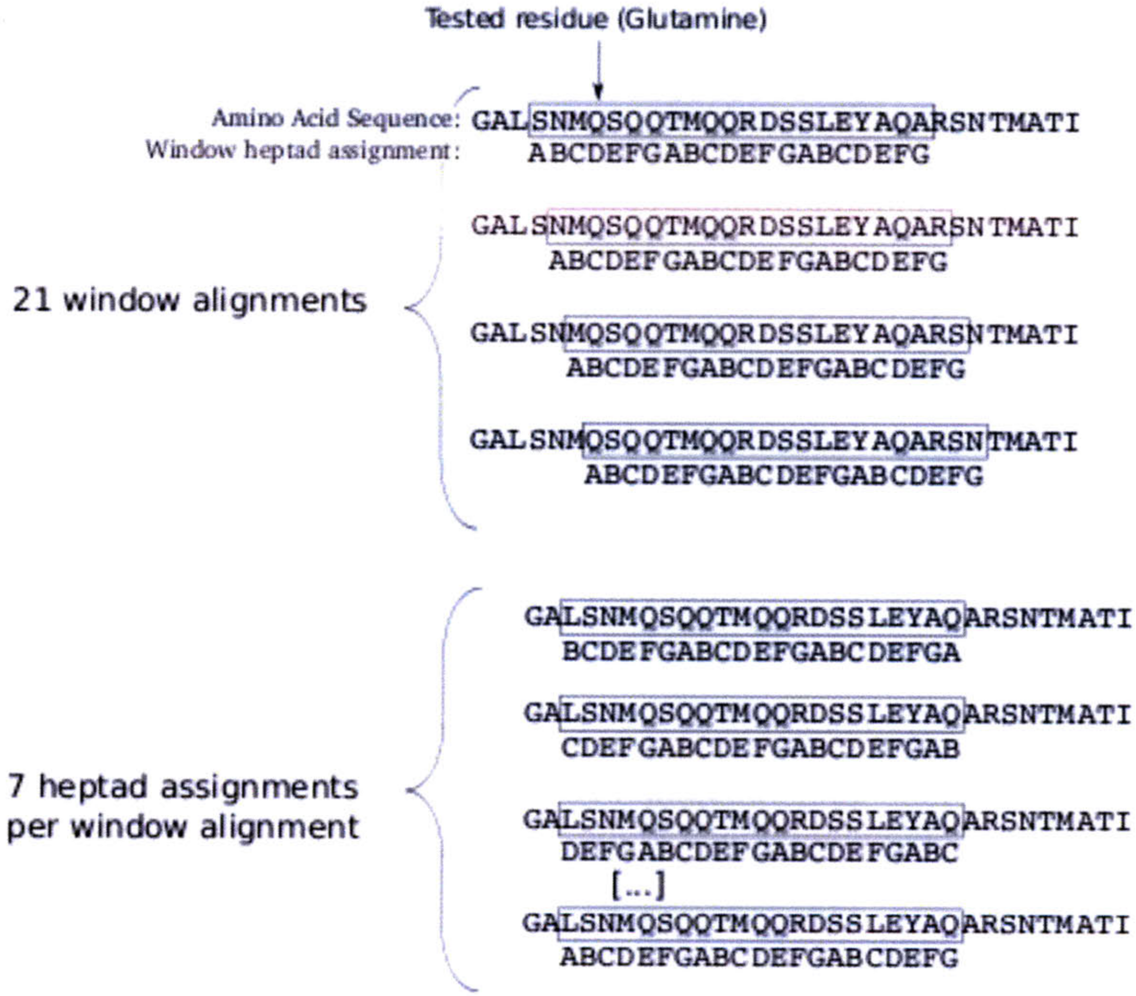
Tested residue (Glutamine)
Amino Acid Sequence: GALSNMQSQQTMQQRDSSLEYAQARSNTMATI
Window heptad assignment: ABCDEFGABCDEFGABCDEFG
GALSNMQSQQTMQQRDSSLEYAQARSNTMATI
ABCDEFGABCDEFGABCDEFG
21 window alignments
GALSNMQSQQTMQQRDSSLEYAQARSNTMATI
ABCDEFGABCDEFGABCDEFG
GALSNMQSQQTMQQRDSSLEYAQARSNTMATI
ABCDEFGABCDEFGABCDEFG
GALSNMQSQQTMQQRDSSLEYAQARSNTMATI
BCDEFGABCDEFGABCDEFGA
GALSNMQSQQTMQQRDSSLEYAQARSNTMATI
CDEFGABCDEFGABCDEFGAB
7 heptad assignments
per window alignment
GALSNMQSQQTMQQRDSSLEYAQARSNTMATI
DEFGABCDEFGABCDEFGABC
[...]
GALSNMQSQQTMQQRDSSLEYAQARSNTMATI
ABCDEFGABCDEFGABCDEFG

Figure 2: Sample PSSM

choosing an optimal scoring method, the PSSM scoring methods have typically been chosen by intuition.

COILS

For example, the first popular algorithm, COILS (Lupas et al 1991), compiled a database of coiled coils. From this database, the authors created a 7x20 table; for each heptad and amino acid, they found the probability of that amino acid occurring at that heptad location. Similarly the authors created a 20 entry table for the frequency of each amino acid in non-coiled coil sequences. From these two tables, the authors scored their window as follows: For each residue i in the window, let $P(i)$ be the probability of the amino acid at location i occurring at the proposed heptad position, and compute $\sum_i \log P(i)$. Also, compute the probability $N(i)$ of that amino acid occurring in proteins from a negative database, and compute the corresponding sum $\sum_i \log N(i)$. The difference between these two statistics is the window score.

Paircoil

The later Paircoil method (Berger et al 1995) noted correlations between nearby residues in coiled coil structures, and incorporated these correlations in its scoring system. In addition to the frequency tables computed by COILS, Paircoil computed 20x20x7 pairwise probability tables for each of several distances k. For distance k, the table included, for each heptad position h and each pair of amino acids a_1 and a_2, the probability of finding a_1 and a_2 at heptad positions h and h+k, respectively.

For two locations i and j in the scoring matrix, let $P(i, j)$ be the probability of the amino acid at location i occurring at the given heptad position *and* the amino acid at j occurring at its given heptad position. Note that this probability will be equal to $P(i)P(j)$ if $P(i)$ and

$P(j)$ are independent. Neglecting edge effects, Paircoil computes:

$$\sum_i \frac{1}{3} \log \frac{P(i,i+1)P(i,i+3)P(i,i+4)}{P(i+1)P(i+3)P(i+4)} \tag{1}$$

As well as the equivalent probabilities over the negative data set:

$$\sum_i \frac{1}{3} \log \frac{N(i,i+1)N(i,i+3)N(i,i+4)}{N(i+1)N(i+3)N(i+4)} \tag{2}$$

Paircoil uses the difference of these two values as its score for the window. Rewriting these propensities:

$$\begin{aligned}
& \sum_i \frac{1}{3} \log \frac{P(i,i+1)P(i,i+3)P(i,i+4)}{P(i+1)P(i+3)P(i+4)} \\
= & \sum_i \frac{1}{3} \log P(i)^3 \frac{P(i,i+1)}{P(i)P(i+1)} \frac{P(i,i+3)}{P(i)P(i+3)} \frac{P(i,i+4)}{P(i)P(i+4)} \\
= & \sum_i \log P(i) + \frac{1}{3} \sum_i \log \frac{P(i,i+1)}{P(i)P(i+1)} + \frac{1}{3} \sum_i \log \frac{P(i,i+3)}{P(i)P(i+3)} \\
+ & \frac{1}{3} \sum_i \log \frac{P(i,i+4)}{P(i)P(i+4)}
\end{aligned}$$

Similarly, non-coiled coil propensity can be written as:

$$\sum_i \log N(i) + \frac{1}{3} \sum_i \log \frac{N(i,i+1)}{N(i)N(i+1)} + \frac{1}{3} \sum_i \log \frac{N(i,i+3)}{N(i)N(i+3)} + \frac{1}{3} \sum_i \log \frac{N(i,i+4)}{N(i)N(i+4)} \tag{3}$$

When written in this form, we can see that the propensities are split into two parts - the first term is the total log probability of the sequence assuming no correlations between residues at different locations, and the last three terms correspond to the correlation between residues 1, 3 and 4 apart (distances the Paircoil authors determined to be the most effective predictors). When viewed this way, there is no obvious reason why the correlation terms should each have equal weightings, or why they should each be given $\frac{1}{3}$ the weight of the log probability term.

We develop a method that will remove these constraints and add more predictors to improve prediction of coiled coil structure.

Multicoil

The Paircoil PSSM method can also be adapted to distinguish between dimer and trimer oligomerization states. Instead of just using the coiled coil probabilities and pairwise probabilities, Multicoil (Wolf et al 1997) computes separate dimer and trimer probabilities $D(i)$ and $T(i)$, along with dimer and trimer pairwise probabilities. The Multicoil method computes 6 PSSM statistics for each residue, which were computed in the standard way by taking the maximum among all windows and heptad alignments. These statistics are the Paircoil statistics for various individual distances.

$$\sum_i \log \frac{D(i,i+2)}{D(i+2)} - \log \frac{N(i,i+2)}{N(i+2)}$$
$$\sum_i \log \frac{D(i,i+3)}{D(i+3)} - \log \frac{N(i,i+3)}{N(i+3)}$$
$$\sum_i \log \frac{D(i,i+4)}{D(i+4)} - \log \frac{N(i,i+4)}{N(i+4)}$$
$$\sum_i \log \frac{T(i,i+3)}{T(i+3)} - \log \frac{N(i,i+3)}{N(i+3)}$$
$$\sum_i \log \frac{T(i,i+4)}{T(i+4)} - \log \frac{N(i,i+4)}{N(i+4)}$$
$$\sum_i \log \frac{T(i,i+5)}{T(i+5)} - \log \frac{N(i,i+5)}{N(i+5)}$$

The method fits three Gaussians (one each for dimer coil, trimer coil and non-coiled coil) to the distribution of these statistics and uses the resulting Gaussians to predict future residues.

HMM methods

In response to the shortcomings of the PSSM methods, HMMs have been used in coiled coil prediction, by Marcoil among others (Delorenzi and Speed 2002; Lisa Bartoli et al 2009). The HMM methods treat the structure at each residue as a hidden state, and the observed amino acid at that location as the value emitted by that state. By defining the emission probabilities of each state and the transition probabilities between states, we can use standard HMM algorithms to predict the probability of each hidden state at each location. We can then predict coiled coil behavior by summing across states that represent the same conformation.

In particular, the Marcoil method uses one state for the non-coiled coil structure and $7 * 9 = 63$ states to represent the coiled coil structure. These 63 states are labelled by heptad position (abcdefg) and also by location of the amino acid within the coiled coil structure (1-9). The states 1-4 represent the first four residues of the coil, state number 5 represents all the middle residues of the coil, and states 6-9 represent the last four residues of the coil. One more state, state 0, represents a state of not being in a coiled coil. Figure 3 shows the possible transitions between states (neglecting the skip probabilities, which the authors claim have little impact on the results), and Figure 4 shows a possible path through states for a sample sequence.

For the states corresponding to coiled coils, probability of emission equals the probability of the amino acid occurring at that heptad position, the same as the $P(i)$ from the PSSM methods. For the non-coiled coil state, emission probability equals the frequency of the amino acid in the negative database, which is the same as $N(i)$ above.

The transition probabilities (again, neglecting the skip probabilities) are given as: The non-coiled coil state 0 has probability c_1 of transitioning to any of the initial coil states 1a,1b,...1g, and the remaining probability, $1 - 7c_1$, of remaining in state 0. States 1-4 all transition to the next greater number with probability 1. State 5 transitions to 6 with

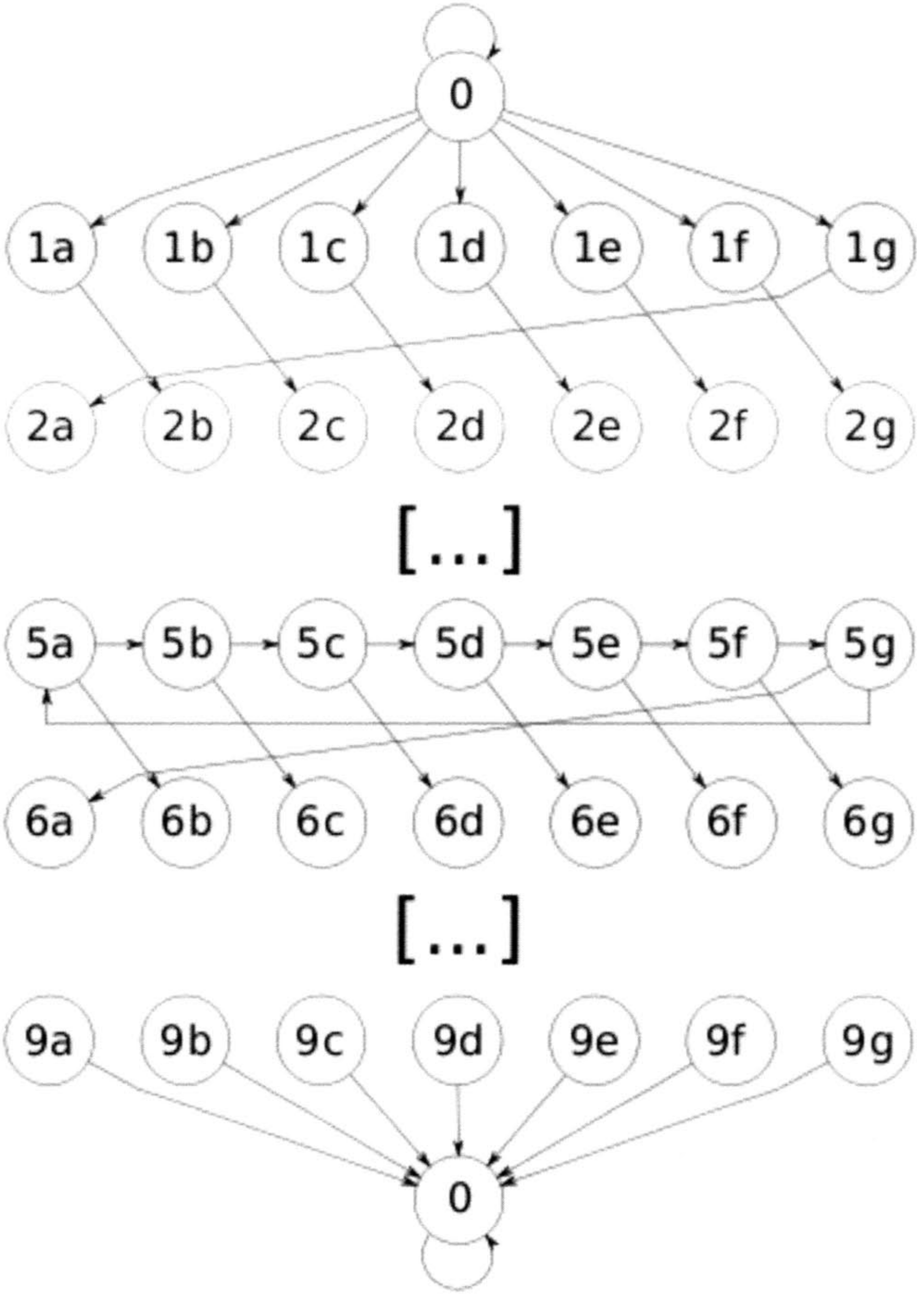

Figure 3: The Marcoil HMM model. States are labelled 1-9 to indicate location in the coiled coil, and a-g to indicate heptad position. Coiled coil states (all states except state 0) transition to states of the next heptad position and next location in the coiled-coil structure, with the exception of position 5, which can also transition to another state of position 5.

Amino Acid sequence: GALSNMQSQQTMQQRDSSLEY

Observed Amino Acid:	G	A	L	S	N	M	Q	S	Q	Q	T	M	Q	Q	R	D	S	S	L	E	Y
Hidden State:	0	0	0	1d	2e	3f	4g	5a	5b	5c	5d	5e	6f	7g	8a	9b	0	0	0	0	0

Figure 4: Example of hidden states corresponding to a given amino acid sequence in the Marcoil HMM model.

probability c_2, and transitions to state 5 again with probability $1-c_2$. States 6-8 all transition to the next number with probability 1, and 9 transitions to state 0, the non-coiled coil state, with probability 1. All states transition to a state with the next heptad position as shown in Figure 3.

For a coiled coil of length l, the probability of a sequence of states corresponding to a coiled coil is:

$$c_1 c_2 (1-c_2)^{l-9} \prod_i P(i) = \frac{c_1 c_2}{(1-c_2)^9}(1-c_2)^l \prod_i P(i) \tag{4}$$

Which is the transition probabilities times the probabilities P(i) of each residue in the coil. On the other hand, the probability of a sequence of states of length l corresponding to a non-coiled coil is:

$$(1-c_1)^l \prod_i N(i) \tag{5}$$

Under this model, there are an exponential number of paths through states. Marcoil uses the HMM forward-backward algorithm to efficiently sum over the probabilities of all possible paths through states to find, at each position, the total probability of the each state at that position. The probability of the non-coiled coil state is the probability that the residue does not lie in the coiled coil, and the sum of the probabilities of the other states is the probability that the residue does lie in a coiled coil. Although the forward-backward algorithm actually involves a summation over possible paths through states, we can to first approximation expect to find coiled coils wherever we can place a coiled coil such that the value in equation 4 is greater than the probability in equation 5.

That is, the HMM method essentially searches through the peptide sequence in linear time for any potential coiled coil with equation 4 much greater than equation 5. The statistics in equations 4 and 5 are similar to those used by COILS, which, when exponentiated, give $\prod_i P(i)$ and $\prod_i N(i)$ (the Marcoil statistics also include extra c^l terms for constant c). Thus we can see that the HMM method is actually quite similar to COILS without the fixed-window size restriction.

The HMM method includes parameters c_1 and c_2 that were chosen by intuition and trial-and-error. We would rather optimize over these parameters. Also, the HMM methods lack the correlation terms from Paircoil. We might consider using higher order HMMs with more complex states to represent the prior few residues of the sequence, but given that we might want to look for correlations as many as 7 residues apart, the number of states and transition probabilities to predict is too large. We need a smaller parameter space to search.

Method

We propose the Multicoil-HMM method which retains the linear runtime of the HMM methods as well as the flexibility of the HMM algorithms for finding coiled-coil sequences of different lengths. At the same time, Multicoil-HMM includes the correlation terms proved valuable by Paircoil. Unlike these current methods, we optimize our parameters for modelling coiled coil propensity.

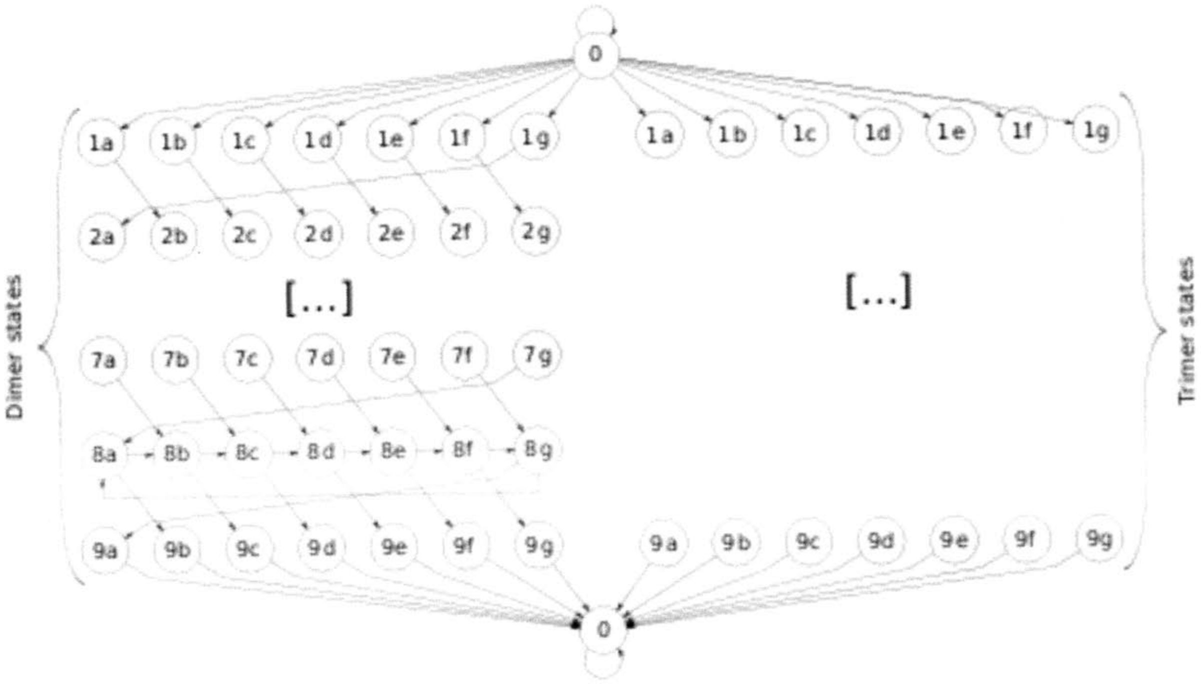

Figure 5: The HMM model. There are two copies of each state, one for the dimer state and one for the trimer state (the transitions of the trimer states are omitted - they are identical to the dimer transitions).

HMM structure

The Multicoil-HMM states are similar to those of Marcoil. There is one non-coiled coil state, $7*9 = 63$ dimer states and $7*9 = 63$ trimer states. The dimer and trimer states are labelled by heptad position (a-g) and by location within the coiled coil (1-9). Slightly differently from Marcoil, the first seven residues of the coil are labelled 1-7, all the middle residues are labelled 8, and the last residue is labelled 9. Transitions are as shown in Figure 5.

Constructing emission probabilities

As noted above, the HMM solution with $P(i)$ and $N(i)$ as its emission probabilities essentially searches through the amino acid sequence, assigning coiled coils at any locations where the statistic for coiled coil propensity

$$\frac{c_1c_2}{(1-c_2)^9}(1-c_2)^l \prod_i P(i) \propto \exp(\sum_i \log P(i)) \tag{6}$$

greatly exceeds the non-coil propensity

$$(1-c_1)^l \prod_i N(i) \propto \exp(\sum_i \log N(i)) \tag{7}$$

That is, where the ratio of the propensities is large:

$$\exp(\sum_i(\log P(i) - \log N(i)) \gg 1 \tag{8}$$

In light of this observation, we seek a set of emission probabilities that correspond to statistics that are better predictors of coiled coil propensity than the simple difference $\sum_i(\log P(i) - \log N(i))$.

Note that in this context there is no reason for emission and transition probabilities to be normalized to 1. Thus we drop the probability interpretation of HMMs and consider unnormalized emission and transition probabilities. Also we set all the transition probabilities to 1 and only vary the emission probabilities. As a result of not normalizing the probabilities, running the forward-backward algorithm on the resulting pseudo-HMM gives likelihoods of each state which no longer sum to 1. After running the algorithm we normalize the likelihoods at each residue to get the probability of each state at the residue.

Also, because the emission probabilities are not normalized, we can scale them such that the non-coiled coil state emission probability is always 1. Instead of taking the coiled coil emission probability as $P(i)$ and the non-coil state emission probability as $N(i)$, we set the non-coil state emission probability to 1 and set the coiled coil emission probability as $\frac{P(i)}{N(i)} = \exp(\log P(i) - \log N(i))$.

Once emission probabilities no longer need to be normalized, there are many possibilities to consider. First of all we can generalize the statistic $\exp(\sum_i(\log P(i) - \log N(i))$ to $\exp(\alpha_1 \sum_i \log P(i) + \alpha_2 \sum_i \log N(i))$ by replacing the emission probability $\exp(\log P(i) - \log N(i))$ with $\exp(\alpha_1 \log P(i) + \alpha_2 \log N(i))$. Now we can try to optimize predictive ability over α_1 and α_2. Before α_1 and α_2 were fixed at 1 and -1 respectively.

Consider adding another predictor: the sum of the hydrophobicities at the a heptad positions of the coil. By choosing as the emission probability $\exp(\alpha_1 \log P(i) + \alpha_2 \log N(i) + \alpha_3 H_a)$ where H_a is the hydrophobicity of the residue for states corresponding to the a heptad, and 0 otherwise, we generate the proper statistic for a path through coiled coil states:

$$\exp(\alpha_1 \sum_i \log P(i) + \alpha_2 \sum_i \log N(i) + \alpha_3 \sum_i H_a)$$

In general, we choose a variety of candidate predictors X_j (such as $\sum \log P(i)$, $\sum \log N(i)$ and $\sum H_a$ above) that we can generate through the proper emission probabilities. Then by raising each of these emission probabilities to the α_j power and multiplying them together we get the total probability of a path through states:

$$\exp(\sum_j \alpha_j X_j)$$

Multicoil-HMM has roughly twice as many states as the Marcoil HMM, half of which correspond to the dimer oligomerization state, and the other half the trimer state. The emission probabilities for the trimer states depend on the same predictors, but with different weights β_i:

$$\exp(\sum_j \beta_j X_j)$$

We need a way to optimize the α_j and β_j, and also to pick good coiled coil predictors X_j which we can represent through pseudo-emission probabilities.

Amino Acid sequence: GALSNMQSQQTMQQRDSSLEY

Observed Amino Acid:	G	A	L	S	N	M	Q	S	Q	Q	T
Hidden State:	0	0	0	1d	2e	3f	4g	6a	7b	8c	8d
	M	Q	Q	R	D	S	S	L	E	Y	
	8e	8f	8g	8a	9b	0	0	0	0	0	

Figure 6: Example of hidden states corresponding to a given amino acid sequence in the Multicoil-HMM model

Including correlations

The above is a significant generalization of the HMM method for coiled coil prediction, but we would also like to include correlation terms which were the main advantage of the Paircoil method. We want the probability of a path through HMM states to depend on correlation predictors; the probability of a path through the coiled coil states should include the correlation term for distance k:

$$X_{corr} = \prod_i \frac{P(i, i+k)}{P(i)P(i+k)} = \exp(\sum_i \log \frac{P(i, i+k)}{P(i)P(i+k)})$$

We achieve this by carefully modifying the emission probabilities of the Multicoil-HMM states. Every state in position at least $k+1$ has its emission probability multiplied by $\exp(\log \frac{P(i,i-k)}{P(i)P(i-k)})$. For an example of how the correlations are included, consider the sequence of hidden states in figure 6.

Consider for example the distance one correlation term. Here the 1d state has its emission probability determined only by single residue predictors. The 2e state's emission probability includes the single residue predictors and is multiplied by a distance 1 correlation term $\exp(\log \frac{P(i,i-1)}{P(i)P(i-1)})$ corresponding to the first two residues of the coiled coil (S and N). The 3f state's emission probability is multiplied by a similar term corresponding to residues 2 and 3 of the coiled coil (N and M), and all successive coiled coil states are multiplied by similar distance 1 correlation terms.

Similarly, the distance two correlation statistic would only affect states of position 3 or greater. The 3f state is further multiplied by a distance 2 correlation term corresponding to residues 1 and 3 of the coiled coil (S and M), the 4g is multiplied by a term corresponding to residues 2 and 4 (N and Q), etc.

Note that because we consider correlations as far as seven residues apart, the emission probabilities might depend on which amino acids are located at the seven prior locations, making it infeasible to precompute the emission probabilities. Instead we modify the HMM procedure to recompute these terms online. This modification increases computation time, but the resulting algorithm still runs in linear time.

Optimizing parameters

These more complicated statistics are only worthwhile if we can pick the right weights. We want, for sequences that represent trimer coiled coils,

$$\exp(\sum_j \beta_j X_j) > \exp(\sum_j \alpha_j X_j)$$

and

$$\exp(\sum_j \beta_j X_j) > 1$$

To ensure that HMM paths through trimer states have more weight than paths through dimer or non-coil states. Similarly, dimer sequences should be such that the dimer probability is greatest, and for non-coil states, the non-coil probability should be greatest. We notice that this type of separation is almost exactly the purpose of multinomial logistic regression, so we can train these parameters by running a multinomial regression on a training set of dimers, trimers and non-coils.

Method details

Coiled Coil Database

The original Multicoil method was trained on a limited set of coiled coils known in 1997. For training and cross-validation we used a new dimer and trimer coiled coil database compiled by Karl Gutwin. The database combined sequences from the Paircoil2 training set, coils from PDB detected by using SOCKET (Walshaw and Woolfson 2001), and other coiled coils from the literature. SOCKET was run on the PQS database as of September 3, 2008 with a distance cutoff of 7.0 Å. The database removed skips and stutters by removing 10 residues on either side of any heptad discontinuity. Sequences less than 21 residues were removed, and the sequences were filtered for 90% sequence similarity by using BLAST alignments on the coiled coil portions of the sequences. The remaining sequences were divided into families by using information from SCOP. For sequences known to not contain coiled coils we used the PDB-minus sequences from Paircoil2 (McDonnell et al 2006).

Regression

As noted above, we use multinomial regression to find the optimal α_j and β_j for the prediction statistic. For our dimer and trimer sequences we used the coiled coil database described above. We also need to train on negative samples; in order to run the regression we need to compute all the candidate predictors on these negative sequences. However, these predictors are dependent on heptad alignment, so we generated sample coils with heptad alignments from the negative PDB-minus database to train on. For each sequence in the PDB-minus negative database, we generated a random integer i uniformly in the range 0-249, generated a random starting heptad h a-g, and took as our coil the first i residues of the negative sequence (or the entire sequence if i was at least the length of the sequence). The coiled coil began at heptad h and continued without any skips.

Then, for each of these dimer, trimer and negative coils, we computed each of the predictors X_j. Given the values of each predictor over each coil, and the correct labelling of each sequence (dimer, trimer, or none), the multinomial regression returns coefficients α_j and β_j such that among the three statistics 1, $\exp(\sum_j \alpha_j X_j)$, and $\exp(\sum_j \beta_j X_j)$, we find 1 is generally the greatest for non-coiled coils, $\exp(\sum_j \alpha_j X_j)$ is the greatest for dimer coils, and $\exp(\sum_j \beta_j X_j)$ is the greatest for trimer coils.

However, the training set included more distinct trimers than dimers, and many more negative coils than either. The regression was conducted in STATA, and we used the pweight option to weight the importance of each element in the training set, normalizing the weight of each sequence in the regression such that the total weight over all dimers and over all trimers were each 1. The total weight over the non-coiled coils is 1000. The value 1000 is arbitrary, but reflects the fact that non-coiled coils are much more common in sequences than coiled coils, and so our priors should strongly prefer predicting non-coiled coil outcomes.

Results

When retrained on this new database of coiled coils and tested using a rigorous leave-family-out training procedure described below, the original Multicoil method correctly identified 64.3% of trimers and 88.1% of dimers. Introducing the Multicoil-HMM algorithm provides a significant improvement over the original Multicoil method. As shown in figure 7 the default setting of Multicoil-HMM produces 70.8% recognition of trimers and 94.0% recognition of dimers. Note that both methods are biased towards dimers. For the Multicoil-HMM method, general coiled coil detection is slightly better than that of Paircoil2, though ac-

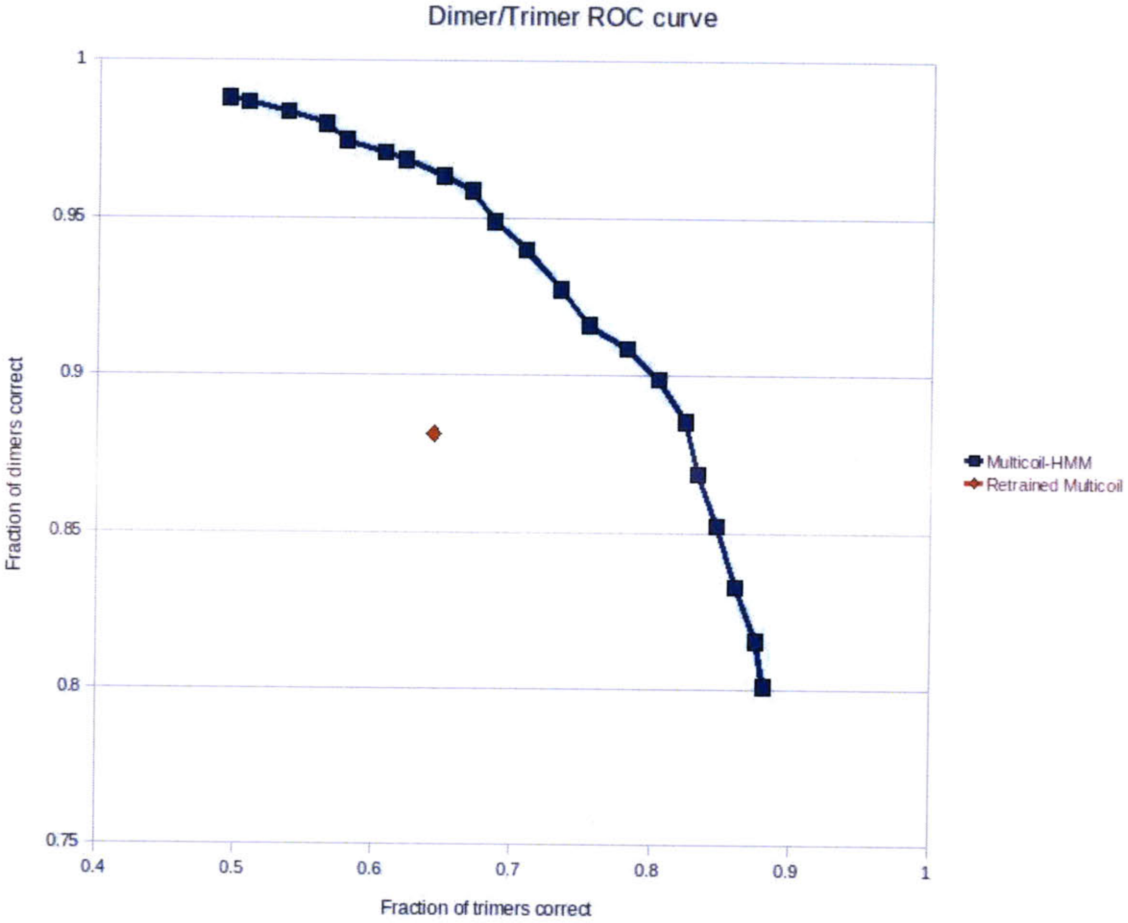

Figure 7: Multicoil-HMM dimer/trimer ROC curve, compared to the retrained Multicoil algorithm. The Multicoil point represents the algorithm's default settings.

tually somewhat worse than a retrained Multicoil algorithm except in the region of high confidence (see figure 8). This test was favorable to the original Multicoil method in that it only includes coils of length 21 or greater. While the issue is disputed (Lupas 1996; Wolf et al 1997), the SOCKET program can identify coiled coils as short as 15 residues long, and the Multicoil-HMM algorithm could be expected to have an additional advantage over the Multicoil algorithm in this realm.

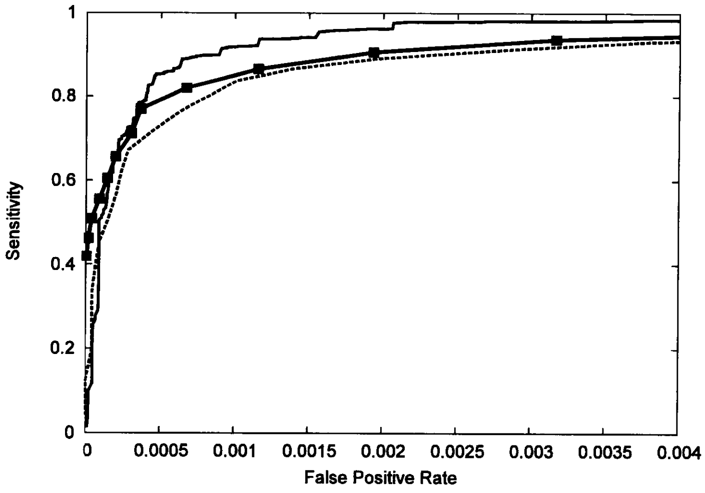

Figure 8: Multicoil-HMM positive/negative ROC curve, compared to Paircoil2 and retrained Multicoil. Multicoil is the solid line, Paircoil2 is the dotted line, and Multicoil-HMM is the solid line with marked points.

Cross-validation

To cross-validate the new Multicoil-HMM method, we used a leave-family-out testing method. For each family in the database, we left out that family and measured performance predicting that family after training on the remaining n-1 families. We used family divisions from the coiled coil database, except we grouped all families with four sequences or fewer into a single miscellaneous family.

However, in order to get optimal results, we also trained the Multicoil-HMM method specially for predicting families that may be different from any of the training families. To train on the remaining n-1 families, each of the n-1 families was left out one at a time, and the frequencies and then predictors X_j for the left-out family were computed based on the data of the other n-2 families. After computing the predictors for each family based only on the data from the n-2 other families, the multinomial logistic regression was fitted to optimally predict a conformation based on predictors.

After the parameters were fitted with this training procedure, the frequencies and predictors were computed based on all n-1 families, and then the conformation of the left-out family was predicted from using these predictors and parameters.

Predictors

We can consider the quality of different regression specifications through the pseudo r-squared value of the regression, and in this way found the best predictors of coiled coil oligomerization state:

- hydrophobicity of residues at the a heptad as measured by the Eisenberg consensus scale (Eisenberg et al 1982).

$$\sum_{heptad=a} H_i$$

- hydrophobicity of residues at the d heptad as measured by the Eisenberg consensus scale.

$$\sum_{heptad=d} H_i$$

- dimer correlations of distances 6 and 7

$$\sum_{k=6,7} \sum_{res} \log \frac{D(i, i+k)}{D(i)D(i+k)}$$

- trimer correlations of distances 1,3 and 6

$$\sum_{k=1,3,6} \sum_{res} \log \frac{T(i, i+k)}{T(i)T(i+k)}$$

- dimer probability of residue at the a,d heptad

$$\sum_{heptad=a,d} \log D(i)$$

- dimer probability of residue at the e,f,g heptad

$$\sum_{heptad=e,f,g} \log D(i)$$

- trimer probability of residue at the b,c heptad

$$\sum_{heptad=b,c} \log T(i)$$

- trimer probability of residue at the e,f,g heptad

$$\sum_{heptad=e,f,g} T(i)$$

- a dummy variable corresponding to whether the amino acid at the g heptad is small (A,C,D,G,N,P,S,T,V)

$$\sum_{heptad=g} i_{small}$$

- non-coiled coil probability of residue

$$\sum_{res} \log N(i)$$

- non-coiled coil correlation at distances 1-7

$$\sum_{k=1-7} \sum_{res} \log \frac{N(i, i+k)}{N(i)N(i+k)}$$

Two parameters were estimated for each of these predictors, for a total of 22 parameters, plus 2 constants. The constants α_0 for dimers and β_0 for trimers were included in our model by multiplying all dimer position 1 emission probabilities by $exp(\alpha_0)$ and all trimer position 1 emission probabilities by $exp(\beta_0)$. We noticed that these constants cause Multicoil-HMM to be heavily biased towards predicting coils. This is likely because for every residue, there are many potential paths through states that predict the residue to be in a coiled-coil conformation, and only one way for the residue to be in a non-coiled coil conformation, so when we sum over all paths, we expect a strong bias towards predicting coiled coils. To compensate for this bias, we subtracted 20 from both the dimer and trimer constants. This value, while somewhat arbitrary, sets a reasonable threshold for coiled coil detection; Multicoil-HMM operates with .2% false positives and 90.6% true positives when evaluating our positive and negative databases residue by residue under the leave-family-out protocol.

Other predictors, including the length of the coil, dummy variables for the charge of the amino acid at different heptad positions, the size of the amino acid at different heptad positions, and some considerations of the amino acid frequencies immediately before and after the coiled coil did not significantly increase the model's ability to predict coiled coil state.

Conclusion

We have noticed the similarities between existing coiled coil prediction algorithms and combined some their strengths to create the Multicoil-HMM algorithm, which does not suffer from the fixed-window size of PSSM methods, is not restricted from using pairwise probabilities like the HMM methods, and uses statistics that are optimized instead of picked from intuition or heuristic, an improvement over either of those methods. The new method significantly improves oligomer-state predictive ability over the previous best algorithm (Multicoil) when that algorithm is retrained on new data. Multicoil-HMM is also more flexible in the length of predicted coiled coils; because our database only includes coiled coils of length at least 21, it provides a conservative estimate of Multicoil-HMM oligomer state prediction quality relative to Multicoil's on conjectured coiled coils as short as 15 residues.

References

Lisa Bartoli, Piero Fariselli, Anders Krogh, and Rita Casadio CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information Bioinformatics 2009 25: 2757-2763.

Bonnie Berger, David B. Wilson, Ethan Wolf, Theodore Tonchev, Mari Milla, and Peter S. Kim, "Predicting Coiled Coils by Use of Pairwise Residue Correlations", Proceedings of the National Academy of Science USA, vol 92, aug 1995, pp. 8259-8263.

Delorenzi M. and SpeedT., 2002. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics, 18(4):617-625, 2002. Abstract

Eisenberg, D., Weiss, R.M., Terwilliger, T.C., and Wilcox, W. 1982. Hydro- phobic moments and protein structure. Faraday Symp. Chem. Soc. 17: 109 120.

Lupas A, van Dyke M. Stock J. 1991. Predicting coiled coils from protein sequences. Science 252:1162-1164. Lupas, A. 1996. Coiled coils : new structures and new functions. Trends Biochem. Sci. 21, 375382

A. V. McDonnell, T. Jiang, A. E. Keating, and B. Berger Paircoil2: improved prediction of coiled coils from sequence Bioinformatics 2006 22: 356-358.

Singh M, Berger B, Kim PS. LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins. J Mol Biol. 1999 Jul 30;290(5):1031-41.

Walshaw J, Woolfson DN. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J Mol Biol. 2001 Apr 13;307(5):1427-50.

Ethan Wolf, Peter S. Kim, and Bonnie Berger, "MultiCoil: A Program for Predicting Two- and Three-Stranded Coiled Coils", Protein Science 6:1179-1189. June 1997.