Prediction of Structurally-Determined Coiled-Coil Domains with Hidden Markov Models

Piero Fariselli[1], Daniele Molinini[1], Rita Casadio[1], and Anders Krogh[2]

[1] Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy
[2] The Bioinformatics Centre, Inst. of Molecular Biology and Physiology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark
piero.fariselli@unibo.it, casadio@alma.unibo.it, krogh@binf.ku.dk.

Abstract. The coiled-coil protein domain is a widespread structural motif known to be involved in a wealth of key interactions in cells and organisms. Coiled-coil recognition and prediction of their location in a protein sequence are important steps for modeling protein structure and function. Nowadays, thanks to the increasing number of experimentally determined protein structures, a significant number of coiled-coil protein domains is available. This enables the development of methods suited to predict the coiled-coil structural motifs starting from the protein sequence. Several methods have been developed to predict classical heptads using manually annotated coiled-coil domains. In this paper we focus on the prediction structurally-determined coiled-coil segments. We introduce a new method based on hidden Markov models that complement the existing methods and outperforms them in the task of locating structurally-defined coiled-coil segments.

Keywords: Protein structure prediction, Hidden Markov models, coiled-coil domains.

1 Introduction

The coiled-coil is a widespread protein structural motif [1] that has been estimated to be present in 5-10% of the sequences emerging from various genome projects [2]. Coiled-coils have a stabilization function and are frequently involved in protein-protein interaction, cell-activities, signaling and other important cellular processes [1].

Coiled-coils comprise two or more alpha-helices wound around each other in regular, symmetrical fashions to produce rope-like structures [3]. In 1953 Francis Crick and Linus Pauling both proposed models for coiled-coil structures, and although Pauling envisaged a broader set of helix periodicity (4/1, 7/2, 18/5, 15/4, 11/3), the Crick's heptad model gained more popularity, probably because he developed a full mathematical description [3]. The sequence bases of these heptad arrangements are repeating patterns of seven residues, which are labelled from *a* to *g*. A general consensus indicates more hydrophobic residues at *a* and *d* positions, which form a hydrophobic stripe on each helix.

S. Hochreiter and R. Wagner (Eds.): BIRD 2007, LNBI 4414, pp. 292–302, 2007.

After 50 years of protein structures determination, we have now structures in the database endowed with the less common periodicities envisaged by Pauling and this enables us to define more general coiled-coil structures. Additionally, we are now in a position where new methods for coiled-coil prediction can be trained on databases containing also structural-derived coiled coil domains.

Coiled-coil segments can be identified in protein structures computationally, particularly with the SOCKET program [2] that was developed to identify general coiled-coil structures. The SOCKET algorithm recognizes the characteristic "knobs-into-holes" side-chain packing of coiled coils, so that it is possible to distinguish coiled-coils from the great majority of helix-helix packing arrangements observed in globular domains. SOCKET is based on the helix-packing structure and therefore coiled-coil domains can be missed in single chains when the coiled-coil is formed with another chain or in half-determined protein structures. Another invaluable source of information is the SCOP classification database [4], in which coiled-coil domains are carefully and manually annotated, and are identified as a specific class (*h* label). In this paper we use both resources to build a reliable coiled-coil protein database in order to train/test our and other prediction methods.

Several programs for predicting coiled-coil regions in protein sequences have been developed so far, and were parameterized on the basis of the heptad module using manual annotations and sequence similarity inference.

Most of them are based on the notion of position specific score matrices (PSSMs), such as COILS [5], PAIRCOIL [6] and MULTICOIL [7]. Also a machine learning approach (MARCOIL) based on a hidden Markov model was previously described [8]. More recently, PAIRCOIL (PAIRCOIL2 [9]) has been improved so as to include new available data including some structurally derived annotations based on the SOCKET program.

When tested on the long and classical coiled-coil domains, the accuracy of all the programs quoted above is remarkably high, but they are less accurate when they predict short or non classical coiled-coil domains as for example the ones identified by the SOCKET [9]. For this reason, in this paper we specifically focus on the task of predicting the location of structurally-annotated coiled coils domains using new hidden Markov models.

2 Method

2.1 The Protein Database

To build our data set structurally annotated coiled-coil domains, we downloaded the SOCKET pre-computed files from the SOCKET web pages. To weed out homologous pairs, the BLASTCLUST program was adopted (from the NCBI BLAST suite) with default parameters and a similarity threshold of 25%. Only one representative structure was kept from each cluster. This gave 138 sequences (SOCKET138). We also extracted all protein domains from PDB that belong to the coiled coil class according to the SCOP classification. The sequences were filtered to decrease similarity with BLASTCLUST as described above, and this gave a set comprising an additional 111 proteins (SCOP111). These 111 proteins are single

representatives of each new cluster generated by BLASTCLUST that did not contain any SOCKET sequence. Our final combined data set (CC249) consists of 249 proteins with a sequence identity of less than 25%.

Furthermore, we ran the BLASTP all-against-all program on CC249 with the low complexity filter turned off. Some 50 protein pairs have sequence identities greater than 25% presumably in low complexity regions, since they were not detected by BLASTCLUST. We included all 249 proteins in our coiled-coil data set; when splitting our set for cross-validation, we made sure that no proteins in the training set had sequence identity greater than 25% with the corresponding test set. As to annotation, we dealt with two different types of files: the files generated by SOCKET and the coiled-coil domains identified in SCOP. Since in this case there is not an explicit indication where the coiled-coil domain starts, we assigned as coiled-coil regions all the helices identified by the DSSP program [10] that fall into a SCOP coiled-coil domain.

A second data set of proteins, not containing coiled-coil domains, was generated using the PAPIA system [11], by removing proteins containing coiled-coil domains. We also checked that no detectable sequence identity with sequences in CC249 were present. The final 'PAPIA' set consists of 2070 protein chains.

Finally for the sake of comparison we used the data set NEWPDB21 (http://paircoil2.csail.mit.edu/supp/new-pdb21.txt) generated for PAIRCOIL2 by McDonnell and coworkers [9]. NEWPDB21 can be regarded as blind set, since contains coiled-coil segments identified only by SOCKET program and not previously recognized using the classical sequence similarity inference and manual annotation.

Data are available at the web page: biocomp.unibo.it/piero/coiled-coils.

2.2 The Hidden Markov Models

The first model we developed and tested was similar to the MARCOIL one (see [8]), and here it is referred to as MChmm. It is endowed with one state modeling the background and 9 groups of 7 states representing the heptad repeats (*a,b,c,d,e,f,g*). All the states of the same repeat type are tied (they share the same emission probability distributions). This constrains the minimal coiled-coil segment length to nine residues. Contrary to the original MARCOIL model, MChmm has explicit begin and end states, which are silent (non-emitting). Our second model (CChmm1) is quite different from MARCOIL and it is depicted in Figure 1. There is one background state (L) and eight coiled-coil states. The model is fully connected and the heptad order is favored by initializing the transition probabilities, so that the probability to follow the heptad order is close to one (0.94) and that of non-heptad transitions is close to zero (0.01). Moreover, we add one more state called H to the coiled coil model. This state accounts for the deviation from the heptad periodicity, as skips, stutters and stammers [3,12]. Finally, in order to take into account different transition probabilities for sequences that contain one and those that have two or more coiled-coil segments, we introduce a third model (CChmm2) shown in Figure 2.

All training phases were performed using the labeled Baum-Welch algorithm [13] while during testing the maximum accuracy decoding [14] was adopted. In the case of CChmm1, the maximum accuracy decoding converges to the posterior-sum algorithm [13].

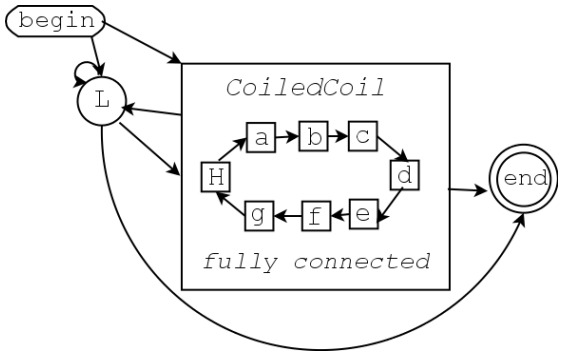

Fig. 1. Automaton representation of the CChmm1 model. The CoiledCoil box represents the coiled-coil states. For sake of clarity only the most probable transitions are indicated.

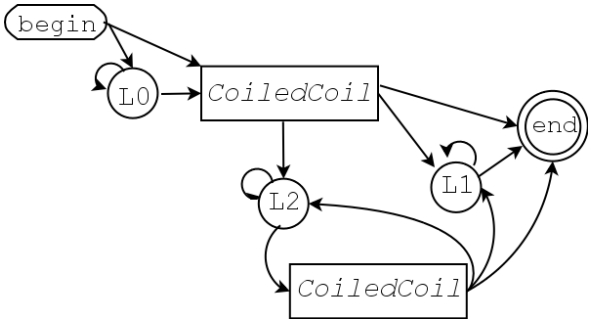

Fig. 2. Automaton representation of the CChmm2 model. The CoiledCoil boxes represent the coiled-coil states as described in Figure 1. The state emission probabilities of two CoiledCoil boxes, as well as those of the L states are tied.

2.3 Scoring the Performance

All the results obtained with our models and other methods are evaluated using the following measures of performance. The fraction of correctly predicted residues is

$$q2 = p/N \tag{1}$$

where p is the total number of correctly predicted residues and N is the total number of residues. This is also used at the sequence level as the fraction of correctly predicted sequences (containing coiled-coil or not), in which case we call it $Q2$. This rule is followed throughout: measures relating to residues are lower case and those relating to complete protein sequences are upper case.

The correlation coefficient for class s is defined as:

$$cor(s) = [p(s)n(s)\text{-}u(s)o(s)\ )] \,/\, d(s) \tag{2}$$

where $d(s)$ is the normalization factor

$$d(s) = [(p(s)+u(s))(p(s)+o(s))(n(s)+u(s))(n(s)+o(s))]^{1/2} \tag{3}$$

For class *s*, *p(s)* and *n(s)* are the numbers of true positive and negative predictions, respectively, and *o(s)* and *u(s)* are the numbers of false positives and negatives, respectively. (Similarly *Cor(s)* and *D(s)* are defined for complete sequences of a given class *s*.)

The coverage or the sensitivity for each class *s* is

$$sn(s) = p(s)/[p(s)+u(s)] \tag{4}$$

The probability of correct predictions (accuracy or specificity) is computed as:

$$sp(s) = p(s) / [p(s) + o(s)] \tag{5}$$

(Similarly *Sn(s)* and *Sp(s)* are defined for complete sequences of a given class *s*.)

In order to score predictions on a segment basis, i.e. to which extent the predicted coiled-coil segments overlap the experimentally determined ones, we compute a segment-overlap measure introduced before [15]. Specifically, we calculate values of the segment overlap accuracy for the coiled-coil regions (SOVC), for the non-coiled-coil region (SOVN). The SOV index is a measure of the intersection divided by the union of the predicted and observed segments [15].

Finally to compute the Receiver Operating Characteristic curve we measured the True Positive Rate that is equal to *Sn(CC)* and the False Positive Rate that is equal to 1- *Sn(N)*.

For comparison, we tested the most recently developed programs MARCOIL, and PAIRCOILS2 using their default parameters. Since MARCOIL give predictions with five different thresholds, we show the best performing threshold. For PAIRCOIL2 we used the decision threshold set to 0.025 and we tested two different size of sliding window (21 or 28 residues, respectively). Since the performances of the two window sizes are almost indistinguishable, as also stated previously by the authors (McDonnell et al., 2006), we show only the best performing one.

3 Results

3.1 Locating Coiled-Coil Segments in Protein Sequences

A major problem in protein structure prediction is the location of coiled-coil regions in proteins. A good prediction of this structural motif can also help in protein modelling procedures. The approaches developed so far (COILS, MULTICOIL, PAIRCOIL, PAIRCOIL2 and MARCOIL), have been proved to be very successful in predicting classical manually annotated coiled-coil domains. However they are less suitable to predict structurally-defined coiled-coil segments [9]. Here we tackle the specific problem of predicting structurally-defined coiled-coil segments (CC249 data set) using different hidden Markov models.

We started developing a HMM similar to that previously described in MARCOIL (MChmm), but using our new structurally-annotated data set CC249. Furthermore we developed and implemented other two HMM models: CChmm1, that does not constrain the structural motif length and CChmm2 that distinguish between chains containing one or more coiled-coil motifs (Figure 1 and 2, respectively). All the

reported results were obtained using a 5-fold cross validation procedure, in which sequence identity between each training and corresponding testing set has less than 25% identity. From Table 1, we can see that the best performing method is CChmm2. This indicates that the CChmm2 model is more suited to capture the information related to the structurally annotated coiled-coils than other HMMs.

Table 1. Performance of different HMM predictors in locating coiled-coil segments in the protein sequence

Method	q2	sn(CC)	sn(N)	SOVCC	SOVN
MChmm	0.75	0.49	0.80	0.52	0.54
CChmm1	0.80	0.57	0.85	0.55	0.63
CChmm2	0.81	0.59	0.86	0.58	0.66

MChmm, CChmm1 and CChmm2 are scored using a 5-fold cross-validation procedure. CC and N represent the coiled-coil class and the non-coiled-coil class respectively.

Table 2. Performance of different HMM predictors in locating coiled-coil segments in the protein sequences of the SOCKET subset

Method	q2	sn(CC)	sn(N)	SOVCC	SOVN
MChmm	0.78	0.38	0.85	0.38	0.64
CChmm1	0.81	0.45	0.87	0.42	0.69
CChmm2	0.81	0.46	0.87	0.43	0.71

For the legend see Table 1.

Table 3. Performance of different HMM predictors in locating coiled-coil segments in the protein sequences of the DSSP-SCOP subset

Method	q2	sn(CC)	sn(N)	SOVCC	SOVN
MChmm	0.69	0.61	0.73	0.64	0.43
CChmm1	0.80	0.71	0.82	0.71	0.56
CChmm2	0.80	0.76	0.84	0.78	0.61

For the legend see Table 1.

Our CC249 training/testing set contains proteins that have been annotated using either SOCKET or DSSP-SCOP. Since the annotation procedure is different for the two methods (see Introduction) this may affect the performance. We therefore evaluated independently the two protein subsets. In Tables 2 and 3 we list the results. The different HMM predictors score similarly, with the exception of MChmm that shows a drop of performance when tested on the DSSP-SCOP subset. One possible explanation is that the DSSP-SCOP subset contains a larger number of short coiled-coil segments that are not easily detected by MChmm. CChmm2 is apparently the best method on both subsets.

3.2 Scoring the Prediction of Different Numbers of Coiled-Coil Segments

CChmm2 was developed to address the problem that the prediction of coiled-coil segments in proteins is usually more difficult for chains containing more than one coiled-coil region. CChmm2 was implemented with different transition probabilities for paths containing one, two or more coiled-coil segments (see Fig. 2). It turns out that this difference is important for the improvement observed for CChmm2. This is apparent from Table 3, where the small increased accuracy due to the protein sequences that contain more than one coiled-coil segment is shown. These findings support our HMM design.

Table 4. CChmm prediction efficiency for the coiled-coil segment location on different subsets

Subset Containing	Method	q2	sn(CC)	sn(N)	SOVCC	SOVN
1 coiled-coil	CChmm1	0.80	0.68	0.88	0.73	0.65
"	CChmm2	0.80	0.68	0.88	0.73	0.65
2 coiled-coils	CChmm1	0.80	0.42	0.86	0.38	0.64
"	CChmm2	0.81	0.43	0.86	0.40	0.65
3 or more	CChmm1	0.76	0.67	0.75	0.53	0.56
"	CChmm2	0.77	0.67	0.75	0.54	0.58

For the legend see Table 1.

3.3 Comparison with Other Methods

The main goal of this work is to develop a predictor of structurally-defined coiled coil regions to complement the exiting predictor in the task of predicting coiled-coil domains starting from the protein sequence. So that is mandatory to compare our CChmm2 with others previously introduced method specifically developed to predict classical heptad coiled-coil domains. We then compare the performance of our CChmm2 with those obtained with the two most recently introduced methods: MARCOIL [8] and PAIRCOIL2 [9]. In Table 5 we report the results of the different predictor on the NEWPDB21 data set (generated by the PAIRCOIL2 authors). This set is based only on SOCKET annotations and can be considered a perfect structurally-annotated blind test. From Table 6 we can see that our CChmm2 outperforms the existing methods on this particular data set, both on residue bases (6 percentage points of q2) and on the overlap between the predicted and observed coiled-coil segments (more than 20 percentage points on SOVCC). This finding indicates that CChmm2 is to be preferred when the prediction focuses on structurally-defined coiled coil segments.

Table 5. Comparison with other methods on the newPDB21(1) data set

Method	q2	sn(CC)	sn(N)	SOVCC	SOVN
MARCOIL	0.70	0.48	0.74	0.46	0.51
PAIRCOIL2	0.71	0.52	0.80	0.48	0.60
CChmm2*	0.77	0.85	0.66	0.73	0.68

newPDB21 is a new blind set previously generated by [9] using SOCKET algorithm. (*) The proteins that showed sequence similarity with those of the training set were predicted using the cross-validation parameters. For the legend see Table 1.

3.4 Discriminating Coiled-Coil Proteins Starting from the Sequence

One of the most important goals in the prediction of protein structure and function is the classification of a protein sequence into a specific structural (functional) class.

It is interesting therefore to evaluate our new implementation in order to discriminating coiled-coil proteins from a set of proteins with different structures, starting from their sequence. This task is very important for structural annotation of whole genomes. The set of proteins containing coiled-coil domains are the true positive examples (CC249) and the filtered PAPIA set contains the negative cases (2070 sequences). To assign a score to each protein sequence with HMMs there are several possibilities. The most natural one is to adopt the probability of the sequence given the HMM model (*P(s|HMM)*). However, the *P(s|HMM)* value is not a good discriminating function, as discussed before [13]. For this reason as a discriminative score for our HMMs (only CChmm2 values are shown), we adopted the posterior probability sum normalized to the protein length. More formally, if $P(\lambda(i)=\Lambda|s)$ is the posterior probability of emitting the *i*-th symbol of sequence *s* in a state whose label is Λ [13], then our score for that sequence is computed as:

$$D(s)=(\Sigma \; P(\lambda(i)=CC|\, s\,)\delta(\, argmax_{\{\Lambda\}}\,(P(\lambda(i)=\Lambda|\, s),CC\,)\,)\,/\,L \qquad (6)$$

where the summation runs over the protein length *L*, δ is the Kronecker delta, Λ is a general label and *CC* is the coiled coil label. This equation gives the sum of the posterior probability labelling for all the positions predicted to be in a coiled-coil state and normalized to the protein length. The score is bounded between zero and 1, since $P(\lambda(i)=CC|\, s\,)$ is always less or equal to 1. In this way, choosing a specific threshold *TH,* a given sequence *s* is assigned to the coiled-coil class when its *D(s)* score is greater than *TH*.

In Figure 3 the ROC curve is obtained with different levels of *D(s)* using the CChmm2 model (the curve for CChmm1 is very similar). From the ROC curve it can be evaluated that CChmm2 scores with a value of Sn(CC) (sensitivity of positive class) equal to 40% when Sn(N) (sensitivity of negative class) is equal to 99%. In this case the error rate is 1% (1- Sn(N)). When a larger error is accepted (35%), sensitivity of the positive class can be as high as 80% (Sn(CC)) (Figure 3).

For comparing with other methods we run the two most recently introduced and best performing predictors (MARCOIL and PAIRCOIL2) on the same testing set comprising both CC249 and the PAPIA sequences for a total of 2319 chains (Table 6). It is worth noticing that the frequency of the coiled-coil proteins in the whole

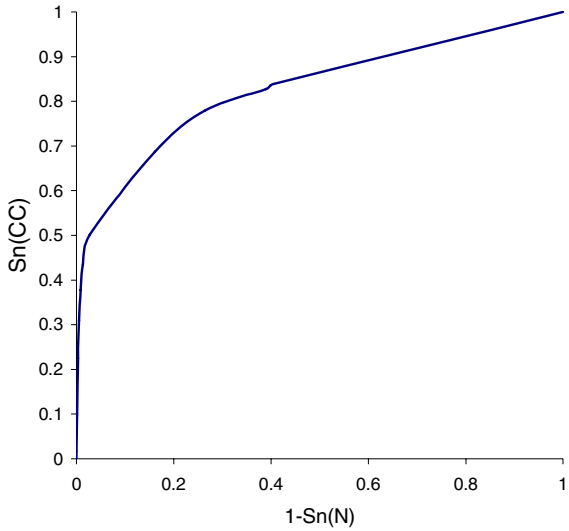

Fig. 3. ROC curve representing the True Positive Rate (Sn(CC)) as function of the False Positive Rate (1-Sn(N)) when CChmm2 is used to discriminate between proteins containing and not containing coiled coil domains. The results are obtained on the non redundant set of globular proteins (PAPIA 2070 proteins) for the False Positive Rate, and in cross validation on the set of 249 coiled coil domains for the corresponding True Positive Rate.

protein set is roughly the same as that estimated in genomes (249/(2070+249) = 0.10; [2]). To compare with the other methods, we report the CChmm2 results using a discriminative threshold set to 0.5 ($D(s) > 0.5$), which was selected to be a reasonable trade-off between the false positive and false negative rates (Fig. 3). For MARCOIL we report the best discriminating threshold that in this case is TH90 (differently from the previous task in which was TH2, see above). All methods are scoring with similar values of Q2 and values of the correlation coefficient ranging from 0.27 up to 0.56 at the most, indicating that the discriminative power for this specific task is not dependent on the coiled-coil annotation type.

Table 6. Discrimination capability of different predictors for coiled-coil-containing proteins using CC259 and PAPIA sets

Method	Q2	Sn(CC)	Sn(N)	Sp(CC)	Sp(N)	Cor
PAIRCOIL2	0.93	0.41	0.99	0.84	0.93	0.55
MARCOIL	0.92	0.40	0.99	0.79	0.93	0.53
CChmm2*	0.92	0.51	0.97	0.69	0.94	0.56

CC and N represent the coiled-coil class and the non-coiled-coil class respectively.
* Present work with a Dcc(s) threshold set to 0.5 (see Eq. 6).

4 Conclusions

In this paper we derive a database of proteins with structurally annotated coiled-coil segments to train and/or test coiled-coil prediction methods. The coiled-coil annotation does not strictly adhere to the original Crick heptad model, but can contain other shorter knob-into-hole helix-packing as detected by SOCKET or assigned by SCOP (probably closer to the original ideas of Pauling [3]). We introduce new HMMs specifically to predict these general types of coiled-coil structural domains, achieving 81% accuracy per residue and a coiled-coil segment overlap of 58%. We also compare our predictor with the two most recent available methods, which have been proved to be very effective in predicting classical coiled-coil domains [8,9] on a SOCKET-derived data set (NEWPDB21) recently introduced [9] and we showed that our method outperform them of 6 percentage points per residue, and 20 percentage points when measured by coiled-coil segment overlap (SOVCC). This indicates that our HMM (CChmm2) outperforms the existing methods in the prediction of structurally-defined coiled-coil domains, so that CChmm2 can complement the existing to predict a broader types of coiled-coil domains.

Acknowledgments. This work was supported by the following grants: MIUR for a PNR-2003 grant delivered to PF. Biosapiens Network of Excellence project (a grant of the European Union's VI Framework Programme, PNR 2001-2003 (FIRB art.8) and PNR 2003 projects (FIRB art.8) on Bioinformatics for Genomics and Proteomics and LIBI-Laboratorio Internazionale di BioInformatica delivered to RC.

References

1. Lupas A., Coiled coils: new structures and new functions. Trends Biochem Sci. (1996) 21, 375-82.
2. Walshaw J, Woolfson DN. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J Mol Biol., (2001) 307,1427-1450.
3. Gruber M, Lupas AN. (2003) Historical review: another 50th anniversary--new periodicities in coiled coils. Trends Biochem Sci., (1998) 28, 679-685.
4. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res, (2004) 32, D226-229.
5. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science, (1991) 252, 1162-1164.
6. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS. Predicting coiled coils by use of pairwise residue correlations. Proc Natl Acad Sci U S A., (1995) 92, 8259-8263.
7. Wolf E, Kim PS, Berger B. MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci. (1997) 6,1179-1189.
8. Delorenzi,M., and Speed,T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics, (2002) 617-625.
9. McDonnell AV, Jiang T, Keating AE, Berger B Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics. (2006) 22:356-358.
10. Kabsch,W. and Sander,C. Dictionary of protein secondary structure: pattern of hydrogen-bonded and geometrical features. Biopolymers. (1983) 22, 2577-2637.

11. Noguchi,T. and Akiyama,Y. "PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003", Nucleic Acids Research, (2003) 31, 492-493.
12. Lupas AN, Gruber M. The structure of alpha-helical coiled coils. Adv Protein Chem. (2005) 70,37-78.
13. Durbin,R., Eddy,S., Krogh,A. and Mitchinson,G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge Univ. Press, Cambridge.
14. Kall L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics. (2005)21, i251-i257.
15. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins, (1999) 34, 220-223.