BIOINFORMATICS

Vol. 18 no. 4 2002
Pages 617–625

An HMM model for coiled-coil domains and a comparison with PSSM-based predictions

Mauro Delorenzi and Terry Speed*

Genetics and Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Victoria 3050, Australia

Received on June 18, 2001; revised on October 17, 2001; accepted on November 15, 2001

ABSTRACT

Motivation: Large-scale sequence data require methods for the automated annotation of protein domains. Many of the predictive methods are based either on a Position Specific Scoring Matrix (PSSM) of fixed length or on a windowless Hidden Markov Model (HMM). The performance of the two approaches is tested for Coiled-Coil Domains (CCDs). The prediction of CCDs is used frequently, and its optimization seems worthwhile.

Results: We have conceived MARCOIL, an HMM for the recognition of proteins with a CCD on a genomic scale. A cross-validated study suggests that MARCOIL improves predictions compared to the traditional PSSM algorithm, especially for some protein families and for short CCDs. The study was designed to reveal differences inherent in the two methods. Potential confounding factors such as differences in the dimension of parameter space and in the parameter values were avoided by using the same amino acid propensities and by keeping the transition probabilities of the HMM constant during cross-validation.

Availability: The prediction program and the databases are available at http://www.wehi.edu.au/bioweb/Mauro/Marcoil

Contact: delorenzi@wehi.edu.au

INTRODUCTION

Coiled-Coil Domains (CCDs) function in the stabilization of the tertiary and quaternary structure of protein molecules. They are frequently involved in protein–protein interactions, and play central roles in diverse processes such as cell-invasion, protein trafficking, signalling and transcription. The experimental verification of binding between proteins predicted to have a CCD is possible on a genomic scale, as has been shown in yeast (Newman *et al.*, 2000). Such approaches are important for unravelling cellular functions, and would benefit from improvements in the identification of putative CCDs. The aim of this study is twofold. On one side we try to optimize the selection of proteins with a CCD in the analysis of genomes. Additionally, we evaluate two different approaches to such problems, one based on a Position Specific Scoring Matrix (PSSM) the other on a Hidden Markov Model (HMM). For reviews about the prediction of CCDs see Lupas (1996a, 1997).

A coiled-coil is formed by the intra-or extra-molecular association of two or more α-helices, which wrap around each other. Each of these single helices is commonly referred to as a CCD. The association is driven by the exclusion of water from the hydrophobic interface (the core). Most CCDs have a 'heptad' repeat, a periodic sequence pattern of seven characteristic residues. The hydrophobic core positions are designated a and d. They are separated by two positions (b and c) respectively by three positions (e, f and g) that are occupied by mainly hydrophilic and often charged residues. The first coiled-coil structures proposed were the long domains of fibrillar proteins (Crick, 1953). In recent years, many cases of shorter CCDs that have important functions have been described. These CCDs often mediate temporary protein–protein interactions and are being found in a growing number of different protein families, for example in the SNARE family and in viral proteins that trigger membrane fusion events (Harbury, 1998; Rothman, 1994; Skehel and Wiley, 1998; Sutton *et al.*, 1998). It is probable, that a number of protein families with CCDs are still to be discovered.

Several programs for predicting CCDs have been described. The most relevant to large-scale annotations are COILS (Lupas *et al.*, 1991), probably the most widely used, PAIRCOIL (Berger *et al.*, 1995) and MULTICOIL (Wolf *et al.*, 1997). Other programs were written for more specific tasks (Bornberg-Bauer *et al.*, 1998; Hirst *et al.*, 1996; Woolfson and Alber, 1995). A search through the literature suggests that COILS, PAIRCOIL and MULTICOIL are considered roughly equally successful in detecting unspecific CCDs.

The approach to finding a CCD by using amino acid propensities was pioneered by Parry (1982). It was perfected and implemented in COILS by A. Lupas and collaborators (Lupas, 1996b; Lupas *et al.*, 1991). The

*To whom correspondence should be addressed.

PSSM stores the seven position specific propensities for the 20 amino acids. Every propensity is given by the (adjusted) ratio of the frequency in a given heptad position to the background frequency of the same amino acid. Two scoring matrices, MTK and MTIDK, are widely used (see COILS documentation at ftp.ebi.ac.uk/pub/software/unix/coils-2.2). In COILS a window of length 28 (four heptads) is applied by default, as this was found to give the best compromise between sensitivity and specificity. Here we call its algorithm PSSM28, irrespective of the matrix used, to distinguish it from the actual COILS program and to indicate the length of the window. PSSM28 is fully determined by eight amino acid frequency distributions. The algorithm computes the moving geometric average of the propensities for seven matrices, each beginning with a different heptad position. Every amino acid receives as score the maximum of the values generated by the 196 combinations of the 7 matrices and the 28 windows in which the amino acid is included (less if it is near an end). The scores are usually mapped onto a number between zero and one, which roughly estimates the probability of the amino acid being in a CCD. For our analysis, this matrix-dependent calibration is irrelevant, since performance is compared at given False Positive (FP) error rates. Shorter window lengths of 14 and 21 are also available, but their higher rates of FP predictions reduce their overall efficiency in the selection of CCDs. Their main application is in the analysis of known or presumptive CCDs.

A possible weakness of COILS is that the positions are modelled as being independent of each other. Berger, Kim and colleagues modified the scoring method through the inclusion of appropriately selected additional factors based on the occurrence of correlated residues. This fixed-length window algorithm is used in PAIRCOIL (Berger *et al.*, 1995), MULTICOIL (Wolf *et al.*, 1997) and LEARNCOIL (Berger and Singh, 1997; Singh *et al.*, 1998, 1999). The correlations appear to be specific to classes of CCDs, in MULTICOIL they contribute to distinguishing between dimeric and trimeric structures. For the general identification of CCDs, or for the identification of new classes of CCDs, the correlations are probably too specific.

HMMs have become a standard technique in sequence analysis (Baldi *et al.*, 1994; Durbin *et al.*, 1998; Eddy, 1995; Hughey and Krogh, 1996; Krogh *et al.*, 1994). HMMs are based on a consistent probabilistic framework for which different good algorithms are known (Baum, 1972; Eddy *et al.*, 1995; Krogh and Riis, 1999; Rabiner, 1989). Their application is not straightforward, since the model's architecture often has to be expressly designed. HMMs are computationally more complex and therefore slower, especially if a large number of state transitions are needed, but they are also more flexible. All approaches with a fixed-length window have two weaknesses in common. When the window is longer than the domain, it also contains neighbouring non coiled-coil residues; when it is shorter, evidence cannot be accumulated.

In the present study we explore whether accuracy can be improved by a window-less HMM, called MARCOIL. To assess the variability of the results and the confidence in the conclusions, a cross-validation approach was taken. In each testing round, the amino acid distributions were derived from a learning set and the same distributions were used in the HMM and in the scoring matrix. A comparison between models with a different number of degrees of freedom can be biased. A richer model can more easily benefit from similarities between the learning and testing sequences, perform better in a test but poorly on unrelated sequences. This is known as the 'overlearning' problem, see for example (Bengio, 1996). To eliminate this source of bias, all the transition probabilities were determined up front and kept constant during the cross-validation.

MODEL: ARCHITECTURE AND PARAMETERS

An amino acid sequence is produced by the hidden Markov chain moving in its state space, and emitting amino acids from the states visited. Each state emits amino acids according to a probability distribution specific to that state.

MARCOIL has 64 states. There is a reference or background state indicated with 0. The other 63 states are denoted by a group number 1–9 and by a letter that refers to the heptad position. Groups 1–4 model the first four residues in a CCD (the N-terminal helical turn), and groups 6–9 the last four (the C-terminal turn); internal coiled-coil residues are from group 5. In the model, a CCD has a minimal length of nine, one residue per group. A sketch for the allowed transitions between groups of states is shown in Figure 1.

Let us give an example. The sequence of two heptads, that start with a b position, requires the following state transitions to occur: 0–1b–2c–3d–4e–5f–5g–5a–5b–5c–5d–6e–7f–8g–9a–0. A characteristic of MARCOIL is that the state chain has to begin with a transition from state 0 to the first state that emits an amino acid. So every CCD, even one that starts with the very first amino acid, has to begin with a transition from state 0 to a state of group 1 and an amino acid emitted from this group. Similarly, the state chain returns to 0 after the last residue was generated, either by state 0 or group 9. Thus, for a protein sequence $\alpha(t)$ of length $n(1 \leqslant t \leqslant n)$, the state chain $\pi(t)$ has length $n + 2(0 \leqslant t \leqslant n + 1)$. With the probability for a transition from state r to state s denoted by $\tau(r, s)$ and the probability for the emission of residue a from state s by $\varepsilon(s, a)$, the joint probability of a given protein sequence α and a given chain of hidden states π can be written as: $P[\alpha, \pi] = \tau(0, \pi(1)) \prod_{t=1}^{n} [\varepsilon(\pi(t), \alpha(t)) \cdot \tau(\pi(t), \pi(t+1))]$.

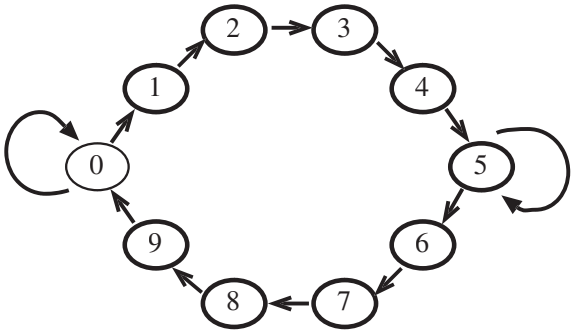

Fig. 1. MARCOIL is an HMM whose states are organized in a background state 0 (thin line) and 9 groups (thick lines) of the seven states that represent a heptad. The figure gives an overview of the transitions that are allowed. A CCD begins with a transition to group 1 and ends with a transition to 0. States of group 5 are revisited for domains with more than 9 residues.

By tying states corresponding to the same heptad position, only eight amino acid distributions were used: $\varepsilon(mx, a) = \varepsilon(1x, a)$ (for every group $1 \leqslant m \leqslant 9$, heptad position x and amino acid a). The set of transition probabilities was parameterized with a triple (i, r, t). In the following equations, let x sweep over the heptad range from a to g, x' characterizes the position that follows x in a heptad and y represents the other six positions (if $x = g$, then $x' = a$ and $y = b, c \ldots g$). With this convention, the equations describe all positive transition probabilities; the symbols u, v and z are merely abbreviations that are used in Figure 2.

(1) $\tau(0, 1x) = i$; $\tau(0, 0) = 1 - 7i$.

(2) $\tau(mx, (m + 1)x') = u = \frac{1}{1+6r}$; $\tau(mx, (m + 1)y) = v = ru = \frac{r}{1+6r} (1 \leqslant m \leqslant 8, m \neq 5)$.

(3a) $\tau(5x, 6x') = t$; $\tau(5x, 6y) = rt$.

(3b) $\tau(5x, 5x') = z = \frac{1}{1+6r} - t$; $\tau(5x, 5y) = rz = \frac{r}{1+6r} - rt$.

(4) $\tau(9x, 0) = 1$ (fixed by the model's structure).

We now explain the role played by the three parameters i, r and t, when the model is used to generate protein sequences.

(1) The probability of the transitions from background to any of the seven states of the first group is the 'initiation probability' i. It controls the frequency of CCDs and it was set to 0.0001.

(2) A constant r is used for the ratio of transition probabilities that do and do not conform to the heptad pattern. It was set to a very low value (0.000 01), so that departures from a perfect heptad pattern are strongly penalized.

(3) The states of group 5 are connected to states of the same and to states of the next group. The transitions internal to group 5 are needed to continue a domain; the transitions to group 6 lead to its last four residues. The parameter t balances between the two options. In the present study two different predictors are used: MARCOIL-L ($t = 0.001$) and MARCOIL-H ($t = 0.01$).

When the model is used, not for generating, but for parsing a sequence, the effect of changing the values of the three parameters is more complex. The frequency and the lengths of predicted CCDs depend mainly on the sequences being processed. A number of preliminary tests indicated a small role for r (as long as it is kept very low). They did show also that the values of i and t act together in determining the lengths of predicted CCDs. Suitable values for r and i were chosen on a trial-and-error basis, testing their role on well-defined CCDs of various lengths (mainly tropomyosins and leucine zippers).

PREDICTIONS: PARSING

Given α, for every position t the conditional probability distributions over the states s is defined by $P[\pi(t) = s|\alpha] = \frac{P[\alpha, \pi(t)=s]}{P[\alpha]}$. These posterior probabilities are calculated with the forward–backward algorithm (Rabiner, 1989). The probability of a residue being in a CCD is taken as the complement of the probability of state 0. The parsing of the sequence is based on the posterior probabilities. If the value is above a threshold, the residue is considered to be part of a CCD. Exactly the same parsing principle is applied to the scores produced by PSSM28.

METHODS AND DATABASES

The performance of coiled-coil predictors is not easily measured, because there is a lack of well-annotated data, especially for emerging types of CCDs. For a proper statistical analysis at the domain level, a reasonable number of sequences and a cross-validation approach were needed. Therefore, CCDs for which there is a clear experimental support in the literature were accepted, even in the absence of confirmation by crystallographic or NMR data. The following underlying assumptions were made:

—performance is given by True Positive (TP) and FP domain predictions;

—a TP is counted for each annotated CCD that overlaps a predicted domain;

—a FP is a predicted domain in a protein sequence known not to contain any CCD.

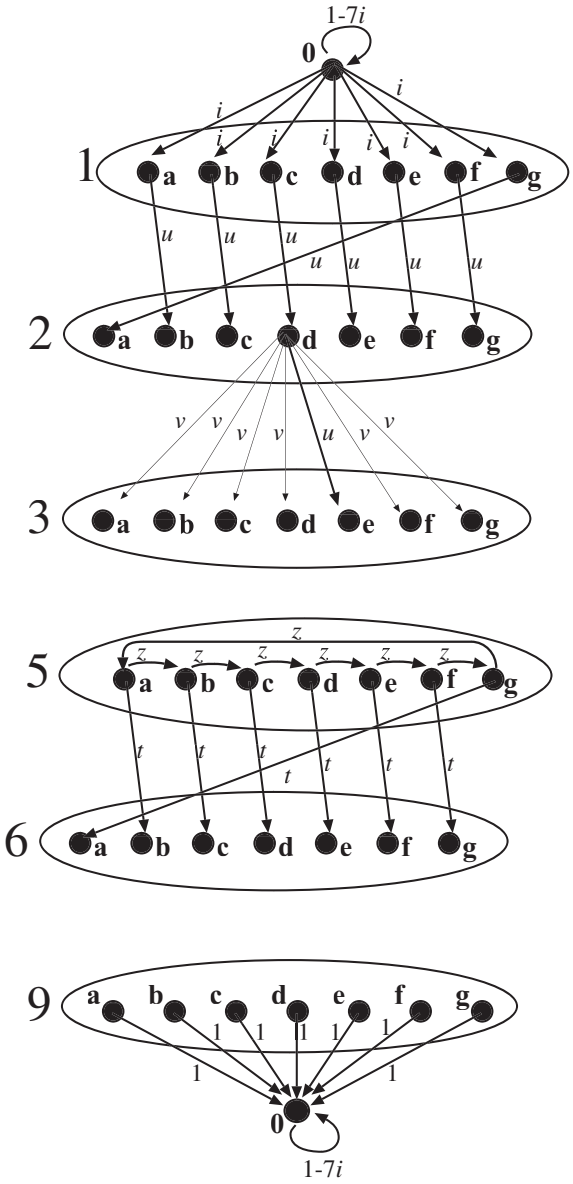

Fig. 2. The model of MARCOIL with some details. State 0 is shown twice for easy representation. Of the possible non-heptad transitions only those starting from state 2d are shown (see text). The groups 1–9 each harbour seven heptad states. Note the special role of group 5 with its internal connections and of group 9, whose states are only connected to state 0. The probabilities are written next to the arrows, and the symbols are defined through the equations in the text.

Moreover, TP and FP counts at the amino acid level tell us something about how well the extent (the position and the length) of the predicted domains agree with the annotations. Some caution is required in interpreting these data, as the real extent of many of the CCDs in the database is not precisely known.

Table 1. The nine classes of protein sequences in the positive database and the numbers of CCDs they have

Section	Name	No. of CCDs
1	Tropomyosins	42
2	Myosins	41
3	Intermediate filaments	202
4	Dyneins	131
5	Kinesins	90
6	Laminins	59
7	SNARE proteins	46
8	Transcription factors	78
9	Other proteins	131

Two databases were used. The sequences in the positive database were derived from the research literature on CCDs and from SWISSPROT (Bairoch and Apweiler, 2000). Since the annotation did not always include all CCDs, the positive test set could not be used for computing FP rates. This database has approximately 100 000 coiled-coil residues of 420 proteins and 820 CCDs that were grouped in nine classes as outlined in Table 1. The class called transcription factors includes members of the bZIP and bHLH families only; the last class collects all proteins that could not be classified in a family of their own, for example human DNA topoisomerase I. The negative database is a collection of 1531 sequences with about 330 000 amino acids devoid of CCDs. It was constructed starting from a collection of entries from the Protein Database PDB (Berman *et al.*, 2000) and eliminating all those that correspond to proteins with known CCDs. Since CCDs were not systematically annotated in PDB, this curation could not be automated and was done manually.

The size and the heterogeneity of this collection of sequences are higher than those used previously for coiled-coil predictions, as we attempted to include the more recently described examples of CCDs. The positive database originally had numerous strong homologies. As explained above, an essential point was to limit the degrees of freedom of the HMM. This should ensure a fair comparison despite homologies. To simulate performance in the recognition of new CCDs, ideally there should be no homologies between the sequences used to estimate the parameters and the sequences used to measure predictive accuracy. However, it can be argued, that new sequences are often related to known sequences, so that a complete elimination of homologies also leads to an unrealistic situation. Practical considerations are also important. A number of sequences sufficient for a significant statistical analysis were desired. Therefore, a very stringent homology reduction was avoided. Pairwise alignments (Smith and Waterman, 1981) were used to

sequentially eliminate sequences, until there was no pair left with over 95% identity in the best local alignment with matrix BLOSUM62 (Henikoff and Henikoff, 1992). As the learning sets are large (280 sequences) and heterogeneous (from nine unrelated families), the effect that a few sequences can have on the scoring of a similar test sequence is limited. Even if homologies might enhance the recognition of some of the CCDs, one would not expect this to substantially affect the comparison.

STUDY DESIGN

Results were obtained from a 150-fold cross-validation test with random partitions of the positive database, two thirds attributed to the learning set, one third to the positive test set. Probability distributions for the seven heptad positions were derived from the learning set. As the location of the CCDs was given, only assignment of heptad positions and counting was required. An automated approach was needed, because the heptad specifications were unavailable in most cases and too numerous for manual classification. As a solution, a modified HMM model and the Baum–Welch algorithm (Baum, 1972) were used for every CCD in the learning set. State 0 was allowed only before and after the last amino acid, so that the coiled-coil residues were assigned to the 63 coiled-coil states. The states representing the same heptad position were tied. The learning runs were initiated with probabilities derived from the MTIDK matrix; the transition parameters used were those of MARCOIL-L and were fixed. Convergence was fast, and three iterations were sufficient to determine the limiting probability values with estimated errors under 10^{-5}. Two modifications of the Baum–Welch algorithm were performed at each iteration. First, the amino acid distributions were determined individually for each of the nine sequence classes, and then averaged. Second, probabilities were smoothed by mixing with 1% background at the end of each iterative step to exclude frequency values of zero. The averaging was chosen to avoid over-representation of the protein families that have a much larger number of coiled-coil residues (the myosins contribute over 30%, the transcription factors about 2%). It thus up-weights the relative contribution of the shorter domains of groups 6–10, whose more difficult prediction we would like to improve. As the smallest class has an average of about 288 residues per heptad position, small sample size errors should not present a problem. The procedure is analogous to the one that had been used for the MTIDK matrix by Lupas *et al.* (see COILS documentation).

Sensitivity was computed at five levels of stringency. For every model the testing consisted of two phases. First, thresholds were computed, so that the number of FP domains in the negative data set was equal to five given FP values. Then they were applied in the analysis of the positive test set. The given numbers of FP domains were set to 105, 68, 53, 34 and 20 respectively. They are the FP numbers of COILS28 with matrix MTIDK and thresholds at 20, 50, 70, 90 and 99% and span the range of stringencies that are usually used.

Table 2. Sensitivity as percentage of the CCDs in the entire test set that was predicted. The numbers represent the average and the standard deviation for the 150 runs in the cross-validation

Level	Predictor		
	PSSM28	MARCOIL-H	MARCOIL-L
1	89.5 ± 1.5	92.0 ± 1.4	92.6 ± 1.4
2	86.7 ± 1.7	88.9 ± 1.6	90.0 ± 1.5
3	84.6 ± 1.7	85.8 ± 1.6	87.6 ± 1.5
4	77.7 ± 2.0	81.7 ± 1.7	83.4 ± 1.5
5	70.1 ± 2.4	76.6 ± 1.8	79.3 ± 1.8

RESULTS

The procedure just described yields a calibration of the MARCOIL predictors in comparison to COILS28. The thresholds fluctuated only by a few percentage points and were on average 27, 46, 60, 78 and 93% for MARCOIL-H and 6, 18, 30, 53 and 88% for MARCOIL-L. The probability scale of MARCOIL-H is thus similar to that of COILS28. The lower t value decreases the coiled-coil probabilities in MARCOIL-L, but they still appear quite realistic (COILS28 is often considered slightly optimistic).

Table 2 shows the results of the 150 fold cross-validation test. About 5% of the domains escape detection already at the first level while about 70% are identified even at high stringency. Comparing predictors, there is an advantage of one to several percentage points for the HMMs at each level. The standard deviations in Table 2 measure how the average changed with the combinations of training and testing sequences. Some of these combinations can be more favourable for correct predictions, so there is a component of variance unrelated to the methods. It is therefore better to use the 150 pairwise differences in TP rates to test the null hypothesis that the differences are centred on zero. Applying a two-tailed t-test the null hypothesis is rejected at every level and for each comparison of two methods. The non-parametric rank sum statistics also indicates that the differences are highly significant. In both tests all the p-values are extremely small (below 10^{-8}). This is due to a high stability of the differences in the cross-validation. Indeed, taking the levels together, MARCOIL-L had a higher number of TP than PSSM28 in all 750 cases, MARCOIL-H in 724 out of 750.

The absolute and relative sensitivity differs between

Table 3. Sensitivity for classes of protein families (average and 1 standard deviation in %). Please refer to Tables 1 and 2

Predictor	Class						
	4	5	6	7	8	9	4–9
a. Level 1							
PSSM28	80.8 ± 5.5	87.9 ± 4.8	98.1 ± 2.5	93.0 ± 6.4	97.7 ± 2.5	68.2 ± 6.9	84.5 ± 2.3
MARCOIL-H	87.6 ± 5.3	89.4 ± 4.9	100.0 ± 0.0	95.8 ± 6.1	100.0 ± 0.0	73.1 ± 6.7	88.3 ± 2.2
MARCOIL-L	87.9 ± 5.3	89.7 ± 4.9	100.0 ± 0.0	95.8 ± 6.1	100.0 ± 0.0	75.5 ± 6.4	89.0 ± 2.1
b. Level 3							
PSSM28	66.7 ± 5.3	81.9 ± 5.2	90.6 ± 5.0	86.6 ± 8.0	96.6 ± 2.9	63.9 ± 8.0	77.3 ± 2.5
MARCOIL-H	64.2 ± 4.9	85.9 ± 4.7	96.7 ± 3.2	88.0 ± 7.7	100.0 ± 0.0	67.1 ± 7.4	79.4 ± 2.3
MARCOIL-L	67.6 ± 5.0	86.0 ± 4.7	98.5 ± 2.3	91.0 ± 7.3	100.0 ± 0.0	70.7 ± 7.3	81.6 ± 2.2
c. Level 5							
PSSM28	37.4 ± 4.5	67.9 ± 8.1	63.8 ± 8.8	46.6 ± 13.0	90.6 ± 4.4	46.6 ± 9.4	56.1 ± 3.4
MARCOIL-H	42.3 ± 3.7	73.1 ± 5.9	89.2 ± 5.1	73.0 ± 8.9	96.7 ± 3.0	50.5 ± 9.4	65.3 ± 2.6
MARCOIL-L	43.6 ± 4.2	77.8 ± 5.7	93.2 ± 4.4	78.2 ± 8.9	98.7 ± 1.9	55.7 ± 9.0	68.9 ± 2.5

Table 4. Sensitivity and specificity as TP and FP rates at the residue level (average and standard deviation in %)

Predictor	Level				
	1	2	3	4	5
a. TP (%)					
PSSM28	90.3 ± 1.1	87.7 ± 1.1	85.5 ± 1.1	79.3 ± 1.4	71.4 ± 1.7
MARCOIL-H	92.9 ± 1.0	90.5 ± 1.1	88.1 ± 1.2	83.7 ± 1.1	77.3 ± 1.3
MARCOIL-L	95.7 ± 1.0	94.7 ± 1.0	93.7 ± 1.1	91.4 ± 1.1	88.2 ± 1.3
b. FP (%)					
PSSM28	1.09 ± 0.01	0.71 ± 0.01	0.56 ± 0.01	0.34 ± 0.00	0.20 ± 0.00
MARCOIL-H	1.05 ± 0.02	0.71 ± 0.02	0.53 ± 0.02	0.33 ± 0.01	0.18 ± 0.01
MARCOIL-L	1.29 ± 0.02	0.96 ± 0.02	0.78 ± 0.02	0.52 ± 0.02	0.30 ± 0.01

protein classes (Table 3). Data are shown for levels 1, 3 and 5 only, as those on levels 2 and 4 are similar. There are no relevant differences between the methods on the first three groups (not shown), whose CCDs are the easiest to detect. The average sensitivity of MARCOIL is always higher on the other six classes, except for the dyneins (class 4). The CCDs of classes 6 and 8, laminins and leucine zippers, are well recognized, while the classes 4 and 9 are the most difficult ones. Sometimes the differences exceed 10 percentage points, indicating that some CCDs are substantially better recognized by the HMM. Again, even where the averages are close, the pairwise differences were quite stable over the cross-validation and significant in a t-test and a rank sum test. For example, in the case of class 5 (kinesins) and level 1, the 150 differences between PSSM28 and MARCOIL-H are incompatible with the null hypothesis in a rank sum test (p-value 10^{-6}). In this same case, the differences between MARCOIL-L and MARCOIL-H are very small and not significant. The last column in Table 3 was obtained with the groups 4–9 pooled.

The two MARCOIL predictors perform similarly on domains, MARCOIL-L marginally higher. They differ more at the amino acid level, as is shown in Table 4. The levels are the same as above, defined by equal numbers of predicted FP domains. MARCOIL-L has higher FP amino acid counts. The observation is easily explained. MARCOIL-L tends to predict longer domains. This is reflected also in its advantage in TP rates being consistently higher at the amino acid than at the domain level (compare with Table 2) and is an expected consequence of a higher value for the t parameter. The performance of PSSM28 can be easily compared to that of MARCOIL-H. MARCOIL-H has consistently higher TP and lower FP rates. The differences between the two are similar in Tables 2 and 4, suggesting that while more domains are predicted by MARCOIL-H, the precision of the predicted domain is similar to that of PSSM28. The trends seen in the class-specific TP rates for amino acids agree perfectly with those seen for domains and are therefore omitted.

Given the differences between protein families, it would be interesting to know more about their causes. As the

Table 5. Sensitivity for domains of different length (average and standard deviation in %)

Predictor	Length				
	1–21	22–28	29–35	36–42	Over 42
a. Level 1					
PSSM28	44.5 ± 1.5	73.0 ± 1.1	96.3 ± 0.2	98.2 ± 0.4	97.8 ± 0.2
MARCOIL-H	58.2 ± 1.4	78.7 ± 1.5	96.3 ± 0.4	98.3 ± 0.0	98.8 ± 0.1
MARCOIL-L	59.7 ± 1.6	80.2 ± 1.4	97.0 ± 0.3	98.3 ± 0.0	99.1 ± 0.0
b. Level 3					
PSSM28	29.5 ± 1.6	57.0 ± 1.5	92.6 ± 0.9	96.2 ± 0.7	96.8 ± 0.2
MARCOIL-H	39.4 ± 2.2	59.8 ± 1.2	89.4 ± 0.6	97.5 ± 0.9	97.7 ± 0.2
MARCOIL-L	46.1 ± 1.5	63.8 ± 1.1	91.4 ± 0.8	98.0 ± 0.6	98.2 ± 0.2
c. Level 5					
PSSM28	11.2 ± 0.8	30.8 ± 1.5	68.8 ± 2.6	75.7 ± 2.6	90.6 ± 0.6
MARCOIL-H	20.1 ± 1.1	42.7 ± 1.4	73.6 ± 0.8	81.2 ± 1.9	95.6 ± 0.2
MARCOIL-L	27.8 ± 1.3	48.2 ± 1.1	75.6 ± 0.8	85.8 ± 1.9	96.7 ± 0.2

same distributions were used, class-specific amino acid preferences are unlikely to be responsible. One likely factor is the length of the domains. Another factor could be the fidelity of the heptad repeat. To study the role of length, domains were classified accordingly and the results are shown in Table 5. Despite the fact that they are up-weighted in the learning phase, the shorter domains are harder to predict. MARCOIL improves their predictions considerably, doubling the sensitivity in some cases. MARCOIL does not improve the predictions of domains of intermediate length, but it does so on the longer domains. Since almost all the long CCDs in the database are well identified by all predictors, this difference becomes apparent only at high stringency.

DISCUSSION

Like other coiled-coil prediction programs, MARCOIL limits the number of FP predictions by penalizing irregularities in the heptad pattern. In the present implementation, the parameter r controls this feature. The efficient recognition of CCDs with deviations from the heptad pattern (Brown *et al.*, 1996; Hicks *et al.*, 1997) remains difficult. The other major source of FP predictions is short fragments with heptad-like patterns. These error sources are restricted through the choice of longer windows in COILS and by lower values for the parameters i and t in MARCOIL. The window-less HMM approach appears to improve the simultaneous recognition of domains of different lengths.

MARCOIL's speed is limited by the forward–backward algorithm. Like the PSSM28 algorithm, it is linear in sequence length and in the number of sequences to be processed, but it requires more operations per residue. The model has 4096 potential transitions, but only 456 are used, requiring about 912 multiplications per residue. On the other hand, after pre-computing the logarithms of the propensities, the calculation of PSSM scores requires 196 additions per residue but no multiplications. Therefore, MARCOIL is much slower. In practice though, the difference is smaller than expected. The relative speed of our implementations of MARCOIL and PSSM28 varies strongly; ratios from 4 to 40 were estimated for realistic applications, depending on the processor and on the amount of output required. MARCOIL spends most time in multiplications, whilst for PSSM28 input/output operations can be speed limiting. To test MARCOIL under the intended scale of use, we generated a list of all predicted CCDs for the 25,123 *Drosophila melanogaster* protein entries in GenBank as at May, 2001. This job required about 20 min on a machine with a dual Pentium III 900 MHz processor. We conclude that our implementation of MARCOIL is suited for large applications, for example the annotation of all known and predicted proteins in the human genome.

MARCOIL could be further developed in a number of ways: optimizing transition parameters, removing ties, adding states to model domains with an endecad rather than a heptad pattern, to name a few. One reason behind the choice of a model with 9 groups was the expectation that it might be helpful to use different ('capping') propensities for the first and last four amino acids of a domain (Lu *et al.*, 1999; Petukhov *et al.*, 1999; Rohl *et al.*, 1996; Sun *et al.*, 2000). Due to the absence of a large set of training sequences with an exact annotation of domain ends, this improvement will have to wait. A tool called SOCKET, which could prove very helpful in setting up such a database, has been recently described (Walshaw and Woolfson, 2001). An open problem in the field, which one could try to approach, is how correlations like those used in PAIRCOIL could be efficiently included

in a HMM, preferably without a large increase in the number of model parameters or in algorithm complexity.

Summarizing, the results indicate that a fixed length window is a notable limitation in the identification of the heptad repeats of CCDs. The study tried to avoid all obvious sources of bias in the comparison to an HMM approach. Thus, the parameter space of the HMM was not fully exploited. This has the advantage of decreasing the risk of overlearning and weak generalization. Further analysis is required to see if higher performance is possible by making full use of its parameter space. Preliminary results indicate that this might not be easily achieved. In general, we believe that HMMs are well suited to model domains which have a variable number of sequence units, as leucine-rich repeats and ankyrin repeats in proteins. MARCOIL has a simple structure that could be used, maybe with some minor changes, to model all such cases. Our implementation of MARCOIL is available and includes a number of output options. A web interface is in preparation.

ACKNOWLEDGEMENT

M.D. was supported in part by Schweizerischer Nationalfondsprojekt 20-50686.97.

REFERENCES

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M.A. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.

Baum,L.E. (1972) An equality and associated maximization technque in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**, 1–8.

Bengio,Y. (1996) *Neural Networks for Speech and Sequence Recognition*. International Thompson Computer Press, Zocatur.

Berger,B. and Singh,M. (1997) An iterative method for improved protein structural motif recognition. *J. Comput. Biol.*, **4**, 261–273.

Berger,B., Wilson,D.B., Wolf,E., Tonchev,T., Milla,M. and Kim,P.S. (1995) Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl Acad. Sci. USA*, **92**, 8259–8263.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bornberg-Bauer,E., Rivals,E. and Vingron,M. (1998) Computational approaches to identify leucine zippers. *Nucleic Acids Res.*, **26**, 2740–2746.

Brown,J.H., Cohen,C. and Parry,D.A. (1996) Heptad breaks in alpha-helical coiled coils: stutters and stammers. *Proteins*, **26**, 134–145.

Crick,F. (1953) The packing of alpha-helices: simple coiled-coils. *Acta Crystallogr.*, **6**, 689–697.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Eddy,S.R. (1995) Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 114–120.

Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.*, **2**, 9–23.

Harbury,P.A. (1998) Springs and zippers: coiled coils in SNARE-mediated membrane fusion. *Structure*, **6**, 1487–1491.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10 915–10 919.

Hicks,M.R., Holberton,D.V., Kowalczyk,C. and Woolfson,D.N. (1997) Coiled-coil assembly by peptides with non-heptad sequence motifs. *Fold Des.*, **2**, 149–158.

Hirst,J.D., Vieth,M., Skolnick,J. and Brooks,C.L. 3rd (1996) Predicting leucine zipper structures from sequence. *Protein Eng.*, **9**, 657–662.

Hughey,R. and Krogh,A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci.*, **12**, 95–107.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Krogh,A. and Riis,S.K. (1999) Hidden neural networks. *Neural Comput.*, **11**, 541–563.

Lu,M., Shu,W., Ji,H., Spek,E., Wang,L. and Kallenbach,N.R. (1999) Helix capping in the GCN4 leucine zipper. *J. Mol. Biol.*, **288**, 743–752.

Lupas,A. (1996a) Coiled coils: new structures and new functions. *Trends Biochem. Sci.*, **21**, 375–382.

Lupas,A. (1996b) Prediction and analysis of coiled-coil structures. *Methods Enzymol.*, **266**, 513–525.

Lupas,A. (1997) Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.*, **7**, 388–393.

Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.

Newman,J.R., Wolf,E. and Kim,P.S. (2000) A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **97**, 13 203–13 208.

Parry,D.A. (1982) Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. *Biosci. Rep.*, **2**, 1017–1024.

Petukhov,M., Uegaki,K., Yumoto,N., Yoshikawa,S. and Serrano,L. (1999) Position dependence of amino acid intrinsic helical propensities II: non-charged polar residues: Ser, Thr, Asn, and Gln. *Protein Sci.*, **8**, 2144–2150.

Rabiner,L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–285.

Rohl,C.A., Chakrabartty,A. and Baldwin,R.L. (1996) Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol. *Protein Sci.*, **5**, 2623–2637.

Rothman,J.E. (1994) Mechanisms of intracellular protein transport. *Nature*, **372**, 55–63.

Singh,M., Berger,B. and Kim,P.S. (1999) LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins. *J. Mol. Biol.*, **290**, 1031–1041.

Singh,M., Berger,B., Kim,P.S., Berger,J.M. and Cochran,A.G. (1998) Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. *Proc. Natl Acad. Sci. USA*, **95**, 2738–2743.

Skehel,J.J. and Wiley,D.C. (1998) Coiled coils in both intracellular vesicle and viral membrane fusion. *Cell*, **95**, 871–874.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Sun,J.K., Penel,S. and Doig,A.J. (2000) Determination of alpha-helix N1 energies after addition of N1, N2, and N3 preferences to helix/coil theory. *Protein Sci.*, **9**, 750–754.

Sutton,R.B., Fasshauer,D., Jahn,R. and Brunger,A.T. (1998) Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. *Nature*, **395**, 347–353.

Walshaw,J. and Woolfson,D.N. (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.*, **307**, 1427–1450.

Wolf,E., Kim,P.S. and Berger,B. (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.*, **6**, 1179–1189.

Woolfson,D.N. and Alber,T. (1995) Predicting oligomerization states of coiled coils. *Protein Sci.*, **4**, 1596–1607.