The coiled-coil is a common protein structural motif (Lupas, 1996) (Conway and Parry, 1990), resulting in a hydrophobic band that positions are hydrophobic (such as leucine, isolecucine and valine) the form: [abcdefg]

The canonical heptad repeat is often represented in

over two helical turns (7/2) with an anti-parallel super-coil of

Crick’s model is characterized by a ‘knobs-into-holes’ packing of

and Linus Pauling proposed the first model of super-coiled helices. Since the heptad repeat is the most informative

constraint (Parry et al., 2008), all the methods have been parameterized on the basis of the heptad model. The first and widely used COILS (Lupas, 1996; Lupas et al., 1991) exploits the residue frequencies computed on the heptads of the experimentally determined structures known at that time. PAIRCOIL (Berger et al., 1995) and the retrained version PAIRCOIL2 (McDonnell et al., 2006) are based on pair-wise residue correlations. MULTICOIL extends PAIRCOIL to the identification of three-stranded coiled-coils (Wolf et al., 1997). All these methods are based on the notion of position specific score matrices (PSSM). Single sequence-based hidden Markov models (HMM) were also developed to address the coiled-coil prediction: MARCOIL (Delorenzi and Speed, 2002) and CCHMM (Fariselli et al., 2007). For coiled-coil motifs contained in viral proteins it was, however, necessary to develop specialized predictors (Singh et al., 1999).

It is very well known that evolutionary information in the form of sequence profile often increases the overall accuracy of a predictive method (Rost and Sander, 1993). So far, only the COILS method was so far modified to exploit evolutionary information. The profile-based version PCOILS (Gruber et al., 2005) substitutes sequence–profile comparisons with profile–profile comparisons. Even though it was not retrained, PCOILS performs better than the original COILS (Gruber et al., 2005).

In this article, we introduce for the first time a HMM that exploits evolutionary information for discriminating coiled-coil sequences and for locating coiled-coil residues within sequences.
The first sequence–profile-based HMM was designed by Martelli et al. (2002) for predicting and discriminating β-barrel membrane proteins. Furthermore, we expand a recent comparative analysis of coiled-coil prediction methods by Gruber et al. (2006) by testing the available methods on a new blind structurally determined dataset and by scoring them on the basis of per-residue, per-segment and per-protein indices. CCHMM_PROF in a discrimination task scores higher than previous implementations with a per-residue accuracy value as high as 0.97 and a Matthew’s correlation coefficient value of 0.67. When CCHMM_PROF is tested on a localization task, its performing values of accuracy and correlation coefficient are: 0.80, 0.61 per-residue; 0.81, 0.64 per-segment; and 0.80, 0.75 per-protein, respectively. These scores are higher than those obtained with other available predictors on the same benchmark. Furthermore, we show that our predictor, in contrast to others, can generalize also on viral proteins.

1.1 The dataset

In order to generate our dataset, we followed the suggestions of Lupas and co-workers (Gruber et al., 2006) and used the intersection between SCOP and SOCKET. Following this procedure, we ended up with a final annotated dataset comprising 104 sequences (S104). The complete S104 dataset consists of 10,724 residues with 3,565 coiled-coil residues (33% of the overall dataset). Sequences shorter than 30 residues, or with coiled-coil domains shorter than 9 residues were excluded. This limit copes with the 9-residue long domains that are the shortest ones classified by MARCOIL (Delorenzi and Speed, 2002). However, the majority of the proteins contain coiled-coil segments longer than 9 residues (90% of S104 contains coiled-coil segments > 15 residue long; 70% of S104 contains coiled-coil segments > 20 residue long). For testing different methods on a blind set, we selected a subset of 50 non-identical protein chains (S50) with sequence identity < 30% with respect to the sequences of the MARCOIL dataset (Delorenzi and Speed, 2002). The S50 dataset contains 4,903 residues, among which 1,696 belong to coiled-coil regions (~35%). The complement of S50 in S104 (namely S104– S50) is labeled as S54. With our procedure no sequence in the S50 dataset has > 30% sequence identity with any of the sequences in the S54 dataset.

To verify that our method was not overfitting data, we performed a cross-validation procedure dividing the S104 dataset into five disjoint sets whose sequence identity is < 30% between any two sequences coming from different sets. These sets were computed with a transitive closure algorithm by defining a graph whose nodes represent the 104 proteins. An edge connects two nodes if and only if the sequence identity between the corresponding protein sequences is 30%. The transitive closure algorithm defines the clusters as the connected components of the graph. Thus, clusters can contain proteins whose pair-wise sequence identity is very low. For instance, given three proteins $p_1$, $p_2$, and $p_3$ and the similarity measure $S$, if the sequence similarities of the three pairs are $S(p_1,p_2) = 30\%$, $S(p_2,p_3) = 30\%$ and $S(p_1,p_3) = 15\%$, the three proteins are assigned to the same cluster (or set) by the algorithm. We also defined a dataset that does not contain coiled-coil domains according to SCOP or SOCKET. We downloaded the Astral SCOP (release 1.69) (Chandonia et al., 2004), which contains sequences with < 40% identity. We further reduced identity to 25% and filtered out all the sequences similar to those contained in the MARCOIL negative set that contains sequences without the coiled-coil motif (Delorenzi and Speed, 2002). The selected sequences were processed with SOCKET (7.4 Å packing cut-off) and all the sequences for which the program detected at least one coiled-coil residue were removed from the dataset. Our negative dataset consists of 1,139 protein sequences (S1139).

The complete dataset of positive and negative examples used to test the method performance in discriminating coiled-coil sequences is S1189 (S1139 + S50). The different datasets used to test the method performance in discriminating coiled-coil sequences is S1189 (S1139 + S50). The different datasets with the cross-validation subsets are available at the web site http://www.biocomp.unibo.it/~lisa/coiled-coils.

For each protein, we computed its sequence profile using the PSI-BLAST outputs (Altschul et al., 1997). PSI-BLAST was run with three rounds of sequence alignment against the Uniref90 database using an E-value of 0.001. Uniref90 is a non-redundant subset of the Uniprot database (The Uniprot Consortium, 2008) that contains no pair of sequences with > 90% sequence identity.

1.2 The HMM (CCHMM_PROF)

Our model is depicted in Figure 1. CCHMM PROF has three background states, labeled with L0, L1 and L2, which model the connections between coiled-coil segments. These three states are tied, which means that they share the same emission probabilities. Moreover, the HMM has two coiled-coil boxes in order to consider different transition probabilities for sequences that contain both one and two or more coiled-coil segments. Each box has a background state H that accounts for the non-heptad coiled-coil periodicities, such as skips, stutters and stammers (Gruber and Lupas, 2003; Lupas and Gruber, 2005). The box includes eight coiled-coil states which are fully connected and whose transition probabilities are initialized so that the heptad order is favored: the probability to follow this order is close to one while the other transitions have a probability close to zero. The states within the two boxes that correspond to the same repeat type are also tied. The same grammar defined by this automaton was previously proposed by Fariselli et al. (2007) (CCHMM).

The major difference here is that the states emit vectors instead of symbols as described in Martelli et al., (2002). CCHMM PROF was chosen for its higher performance (Fariselli et al., 2007). This model allows to extend the concept of coiled-coil patterns to include the characteristic features of the protein sequences, which is essential for the prediction of coiled-coil sequences.
is the first HMM model of coiled-coils that exploits the sequence profile as encoding input. The HMM model is trained with a labeled Baum-Welch algorithm (Durbin et al., 1998) and its accuracy is tested with the posterior-Viterbi decoding (Fariselli et al., 2005).

1.3 Scoring the performance

The results of the different methods were evaluated using the following definitions. The overall accuracy ($Q_2$), namely the number of correctly predicted residues is

$$Q_2 = \frac{p}{N}$$

where $p$ is the number of correctly predicted residues and $N$ is the total number of residues.

The correlation coefficient ($C$) for a given class $i$ is defined as:

$$C(i) = \frac{p(s(i)+o(s(i)))-d(s(i))}{d(s(i))}$$

where $d(s)$ is the normalization factor.

$$d(s) = \sqrt{\left[\frac{(p(s)+o(s))(n(s)+o(s))}{n(s)+o(s)+u(s)+a(s)}\right]}$$

where $p(s)$ and $n(s)$ are the true positive and true negative predictions for class $s$, while $o(s)$ and $u(s)$ are the numbers of false positive and false negative predictions.

The sensitivity ($Sn$) for each class $i$ is defined as:

$$Sn(i) = \frac{p(s(i))}{p(s(i))+a(s(i))}$$

and it accounts for the coverage of the prediction for each class, positive and negative.

The specificity ($Sp$) is the probability of correct predictions and it is defined as follows:

$$Sp(i) = \frac{p(s(i))}{p(s(i))+o(s(i))}$$

All the scores are averaged over each protein sequence.

So far, coiled-coil predictors were evaluated using the above per-residue indices, such as sensitivity and specificity. However, the per-segment index SOV (Segment Overlap) was defined both to account for the different segment distributions and to evaluate secondary structure segments rather than individual residues (Zemla et al., 1999). In particular, we computed the SOV accuracy both for the coiled-coil regions (SOV(CC)) and for the non-coiled-coil regions (SOV(N)). Following Zemla et al., (1999), SOV is defined as:

$$SOV(i) = 100 \times \frac{1}{N_i} \sum_{S(i)} \left[ \frac{\min \{\min(s_1, s_2)+\delta(s_1, s_2)\} - \max \{s_1, s_2\} \times \text{len}(s_1)}{\max \{s_1, s_2\}} \right]$$

where for a class $i$ (coiled-coil or non coiled coil), $N_i$ is the normalization factor, $S(i)$ runs over the pairs of overlapping segments, $\text{len}(s_1)$ is the length of the observed segment $s_1$, $\min(s_1, s_2)$ and $\max(s_1, s_2)$ are the intersection and the union of the two segments $s_1$ and $s_2$, respectively. And $\delta(s_1, s_2)$ is the correction factor for less penalizing the difference between predicted and observed segment termini. According to Zemla et al., (1999), $\delta(s_1, s_2)$ is computed as the minimal value among the following four:

$\max(s_1, s_2), \min(s_1, s_2), \text{len}(s_1) \times 2$ and $\text{len}(s_2) \times 2$.

To further assess the method performance, we introduced a more stringent measure based on Protein OVerlap (POV). For a protein sequence, POV is a binary measure that can be only 0 or 1. POV is equal to 1 only if the number of predicted coiled-coil segments ($N_p$) is equal to the number observed of coiled-coil segments ($N_o$) and if all the corresponding pairs have a minimum SOV. More formally, for each protein, if the number of predicted coiled-coil segments and the number of observed coiled-coil ones are different, the prediction $P$ is considered wrong:

$$\text{if } N_p \neq N_o \rightarrow P = 0$$

Otherwise, if the number of predicted and of observed coiled-coil segments are the same and if the intersection between the two corresponding segments (namely the predicted segment $p_i$ with the corresponding observed segment $o_i$, for $i = 1, \ldots, N_p = N_o$) is above a fixed threshold, a prediction $P$ is considered as a correct prediction:

$$\text{if } N_p = N_o \text{ and } p_i \cap o_i \geq \text{th, } \forall i \rightarrow P = 1$$

If the number of observed and predicted segments is the same, but at least one pair of corresponding segments has an overlap that is not above the threshold, the prediction is considered wrong. Equation (8) is the definition of our POV index.

We defined two thresholds. The first one is half of the minimum of the segment lengths:

$$\text{th} = \min \left( \frac{L_p}{2}, \frac{L_o}{2} \right)$$

where $L_p$ and $L_o$ are the length of the predicted coiled-coil segment and the length of the corresponding observed segment, respectively.

The second threshold imposes a stricter constraint and is defined as half the mean length:

$$\text{th} = \frac{L_p/2 + L_o/2}{2}$$

For a set of proteins, we compute the average of all $P_i$ (Equation 8) over the total number of proteins $N$ as:

$$\text{POV} = \frac{\sum_i P_i}{N}$$

The final scores are obtained by averaging the values computed for each protein over the whole set. This procedure is more stringent than summing up all the predictions and computing the averaged indices at the end. For this reason, it may happen that both $Sn$ and $Sp$ can be lower than the corresponding $Q_2$.

2 RESULTS

2.1 Discriminating coiled-coil sequences

In protein annotation, an important problem is the structural classification of protein sequences. Reliable methods for automatic coiled-coil annotation are therefore necessary. CCHMM_PROF is trained on coiled-coil segments using evolutionary information (profiles). It is tested on a blind set that contains both coiled-coil proteins (never seen before) and proteins that do not have coiled-coil segments. In this discrimination task, proteins were classified into two classes: (i) proteins containing coiled-coils; (ii) proteins not containing coiled-coils.

The performance of different classifiers was compared by computing the receiver operating characteristic (ROC) curve (Fig. 2), where the true positive rate [namely $Sn(CC)$] is plotted as a
The results indicate that CCHMM_PROF is best performing (with a highest AUC value equal to 0.97). We also tested a 28-residue window, which did not affect the final results of CCHMM_PROF with CHMM and PCOILS with COILS). The values for the highest correlation coefficient (0.61). Also the sensitivity (98%) and specificity (73%) values for the coiled-coil class significantly raise with respect to other methods. Furthermore, the global indices referring to the best per-segment and per-protein efficiencies (81% and 80%, respectively) are about 10 percentage points higher than the best ones reached by the methods developed so far.

The first two rows of Table 3 report the performances of CCHMM_PROF and CHMM when a 5-fold cross-validation procedure on the S104 dataset is adopted. The results further confirm the highest performance of CCHMM_PROF. Finally, considering the results of the method obtained after cross validation (first row) and after training on S50 and testing on S50 (third row), it can be concluded that the improvement of the method is robust and does not suffer from data overfitting.

The ROC curve plots the true positive rate [Sn(CC)] as a function of the false positive rate [1−Sn(N)] when different methods are adopted to recognize proteins containing coiled-coil domains. The false positive rate is computed over the S1139 dataset of negative examples while the true positive rate is computed over the S50 dataset. Methods are described in the legend of Table 1. CCHMM_PROF was trained on the complementary set S54.

The ROC curves in Figure 2 were computed over the S1189 dataset. The values for CCHMM_PROF are computed after training the method on the S54 dataset, which has <30% sequence identity with the MARCOIL dataset (that we used for testing). PAIRCOIL2 and MULTICOIL achieve the same low level of false positive rate (1%) with true positive rates of 47% and 49% respectively (Table 1). A significant improvement is registered with CCHMM_PROF (the corresponding true positive rate value is 79%). We also evaluated the area under the ROC curve (AUC) for the different classifiers with the aim of computing a single scalar value for representing the performance (Table 2). Since a random classifier fluctuates around the diagonal line in a ROC plot, any reasonable classifier should have an AUC value >0.5. From the results, it is also evident that sequence profile input sequence profile when used as input helps in improving the performance (compare results of CCHMM_PROF with CHMM and PCOILS with COILS).

The results indicate that CCHMM_PROF is best performing (with the highest AUC value equal to 0.97).

<table>
<thead>
<tr>
<th>Method</th>
<th>Qn</th>
<th>Sn(CC)</th>
<th>Sn(N)</th>
<th>Sp(CC)</th>
<th>Sp(N)</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCHMM_PROF</td>
<td>0.97</td>
<td>0.79</td>
<td>0.99</td>
<td>0.64</td>
<td>0.99</td>
<td>0.67</td>
</tr>
<tr>
<td>CCHMM</td>
<td>0.96</td>
<td>0.79</td>
<td>0.96</td>
<td>0.46</td>
<td>0.99</td>
<td>0.57</td>
</tr>
<tr>
<td>PCOILS</td>
<td>0.96</td>
<td>0.75</td>
<td>0.96</td>
<td>0.48</td>
<td>0.99</td>
<td>0.57</td>
</tr>
<tr>
<td>COILS2</td>
<td>0.96</td>
<td>0.45</td>
<td>0.98</td>
<td>0.49</td>
<td>0.98</td>
<td>0.45</td>
</tr>
<tr>
<td>MARCOIL</td>
<td>0.96</td>
<td>0.51</td>
<td>0.98</td>
<td>0.59</td>
<td>0.98</td>
<td>0.53</td>
</tr>
<tr>
<td>PAIRCOIL2</td>
<td>0.97</td>
<td>0.43</td>
<td>0.99</td>
<td>0.86</td>
<td>0.98</td>
<td>0.62</td>
</tr>
<tr>
<td>MULTICOIL</td>
<td>0.97</td>
<td>0.49</td>
<td>0.99</td>
<td>0.76</td>
<td>0.98</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Table 1. Performance of different predictors in the coiled-coil discrimination task

2.2 Locating coiled-coil regions in protein sequences and comparative evaluation

Given the relevance of the coiled-coil structural motifs in proteins responsible of a number of biological processes, the quality of the prediction when locating coiled-coil segments in a protein sequence is also an issue. In Table 3, we benchmark the results of CCHMM_PROF in a localization task with the results of other available methods. All the methods were tested on the S50 dataset using their default thresholds. PAIRCOIL2 was scored with the decision threshold set to 0.025 and with a window of 21 residues. We also tested a 28-residue window, which did not affect the final scores (data not shown). It is worth noticing that CCHMM_PROF outperforms other HMM-based predictors and all the other methods. CCHMM_PROF achieves the best per-residue accuracy (80%) and the highest correlation coefficient (0.61). Also the sensitivity (98%) and the specificity (73%) values for the coiled-coil class significantly raise with respect to other methods. Furthermore, the global indices referring to the best per-segment and per-protein efficiencies (81% and 80%, respectively) are about 10 percentage points higher than the best ones reached by the methods developed so far.

Given the relevance of the coiled-coil structural motifs in proteins responsible of a number of biological processes, the quality of the prediction when locating coiled-coil segments in a protein sequence is also an issue. In Table 3, we benchmark the results of CCHMM_PROF in a localization task with the results of other available methods. All the methods were tested on the S50 dataset using their default thresholds. PAIRCOIL2 was scored with the decision threshold set to 0.025 and with a window of 21 residues. We also tested a 28-residue window, which did not affect the final scores (data not shown). It is worth noticing that CCHMM_PROF outperforms other HMM-based predictors and all the other methods. CCHMM_PROF achieves the best per-residue accuracy (80%) and the highest correlation coefficient (0.61). Also the sensitivity (98%) and the specificity (73%) values for the coiled-coil class significantly raise with respect to other methods. Furthermore, the global indices referring to the best per-segment and per-protein efficiencies (81% and 80%, respectively) are about 10 percentage points higher than the best ones reached by the methods developed so far.

The first two rows of Table 3 report the performances of CCHMM_PROF and CHMM when a 5-fold cross-validation procedure on the S104 dataset is adopted. The results further confirm the highest performance of CCHMM_PROF. Finally, considering the results of the method obtained after cross validation (first row) and after training on S50 and testing on S50 (third row), it can be concluded that the improvement of the method is robust and does not suffer from data overfitting.

2.3 Assessment of the performance on proteins whose shorter coiled-coil segments are 20 residue long

Most of the methods were designed to predict long coiled-coil segments. For this reason, we compare the different methods on a subset of proteins whose coiled-coil segments are ≥20 residues (Table 4). In Table 4, the SOV(CC) and POV performances of all
Table 4. Comparative evaluation of the methods in the location of the coiled-coil regions

<table>
<thead>
<tr>
<th>Method</th>
<th>Per residue</th>
<th>Per segment</th>
<th>Per protein</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;Q2&gt;</td>
<td>&lt;Sn(CC)&gt;</td>
<td>&lt;Sn(N)&gt;</td>
</tr>
<tr>
<td>CCHMM_PROF</td>
<td>0.80</td>
<td>0.96</td>
<td>0.64</td>
</tr>
<tr>
<td>CCHMM</td>
<td>0.80</td>
<td>0.74</td>
<td>0.82</td>
</tr>
<tr>
<td>CCHMM PROF</td>
<td>0.80</td>
<td>0.98</td>
<td>0.60</td>
</tr>
<tr>
<td>CCHMM</td>
<td>0.78</td>
<td>0.66</td>
<td>0.78</td>
</tr>
<tr>
<td>PCOILS</td>
<td>0.70</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td>COILS</td>
<td>0.64</td>
<td>0.41</td>
<td>0.85</td>
</tr>
<tr>
<td>MARCOIL</td>
<td>0.70</td>
<td>0.60</td>
<td>0.72</td>
</tr>
<tr>
<td>PAIRCOIL2</td>
<td>0.67</td>
<td>0.55</td>
<td>0.61</td>
</tr>
<tr>
<td>MULTICOIL</td>
<td>0.61</td>
<td>0.41</td>
<td>0.77</td>
</tr>
</tbody>
</table>

The methods are scored on the S50 dataset and the results are computed adopting the default decision thresholds. All the computed indices are averaged over the protein dataset.

Table 4. Comparative evaluation of the methods on proteins whose shorter coiled-coil segments are 20-residue long

<table>
<thead>
<tr>
<th>Method</th>
<th>Per residue</th>
<th>&lt;Q2&gt;</th>
<th>&lt;C&gt;</th>
<th>CC</th>
<th>N</th>
<th>Means</th>
<th>Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCHMM_PROF</td>
<td>0.86</td>
<td>0.69</td>
<td>0.94</td>
<td>0.70</td>
<td>0.97</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>CCHMM</td>
<td>0.78</td>
<td>0.53</td>
<td>0.82</td>
<td>0.60</td>
<td>0.89</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>PCOILS</td>
<td>0.78</td>
<td>0.52</td>
<td>0.85</td>
<td>0.64</td>
<td>0.83</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>COILS</td>
<td>0.69</td>
<td>0.35</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>MARCOIL</td>
<td>0.77</td>
<td>0.50</td>
<td>0.80</td>
<td>0.61</td>
<td>0.82</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>PAIRCOIL2</td>
<td>0.72</td>
<td>0.18</td>
<td>0.63</td>
<td>0.43</td>
<td>0.64</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>MULTICOIL</td>
<td>0.65</td>
<td>0.23</td>
<td>0.50</td>
<td>0.51</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

The methods are scored on the subsets of S104 (72 proteins) that contains chains whose shorter coiled-coil segments are ≥20 residue long. CCHMM_PROF and PCOILS use as input sequence profile.

The effect of evolutionary information is first highlighted by comparing PCOILS and COILS (Fig. 3; Table 4), being the only difference between the two methods the sequence-profile input information versus single sequence. The improvement introduced by sequence profile is also evident by comparing CCHMM and CCHMM PROF that overpasses all the methods rather independently of the coiled-coil segment length up to 50 residues (Fig. 3). For coiled-coil segments with a length ≥50 residues all the methods reach a good performance, with the exception of CCHMM.

2.4 Biological insight: targeting viruses

All the correct predictions of the different methods tested on S104 are also correctly assigned by CCHMM_PROF. Given its highest performance, we were able to select a subset of coiled-coil structures correctly predicted only by CCHMM_PROF. Among them, there are interesting targeting virus proteins, such as: the Human Respiratory...
Table 5. CCHMM_PROF correctly predicts coiled-coil regions in viral proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>Uniprot annotationa</th>
<th>Socket annotationb</th>
<th>CCHMM_PROF predictionc</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRSV</td>
<td>155–183</td>
<td>167–199</td>
<td>164–200</td>
</tr>
<tr>
<td>(lg2A,P11209)</td>
<td>558–603</td>
<td>558–603</td>
<td>(647–665)</td>
</tr>
<tr>
<td>SIV</td>
<td>(2ezoA, Q88018)</td>
<td>(644–672)</td>
<td>(616–626)</td>
</tr>
<tr>
<td>Ebola gp2</td>
<td>554–595</td>
<td>561–593</td>
<td>558–593</td>
</tr>
<tr>
<td>(2eboA, Q05320)</td>
<td>(615–634)</td>
<td>(616–626)</td>
<td>(–)</td>
</tr>
</tbody>
</table>

aProtein: for each protein the PDB ID and Uniprot accession code are indicated. (HRSV fusion protein core, the gp41 ectodomain of SIV and the core structure of gp2 from Ebola virus).
bUniprot annotation: the position of the coiled-coil segment in the protein chain as listed in Uniprot 56.0.
cSOCKET annotation: for SOCKET prediction details see Walshaw and Woolfson (2001).

dCCHMM_PROF prediction: our method at work.

Synctial Virus (HRSV) protein fusion core, the ectodomain of the Simian Immunodeficiency Virus (SIV) glycoprotein 41 and the core of glycoprotein 2 from Ebola Virus.

Viral fusion or transmembrane glycoproteins are the major responsible of the entry of a virus in the host cell. It is known that most of these structures are anti-parallel trimeric coiled-coil heterodimers (Caffrey et al., 1998; Lu et al., 1995; Malashkevich et al., 1998, 1999; Matthews et al., 2000).

In Table 5, the annotated and predicted coiled-coil segments (with CCHMM_PROF) in HRSV fusion protein core (PDB id 1g2c, chain A), in the gp41 ectodomain of SIV (PDB id 2ezo, chain A) and in the core structure of gp2 from Ebola virus (PDB id 2ebo, chain A) are reported. For each structure, the Uniprot annotation, if available, and the annotation provided by the SOCKET program are significantly overlapping (overlap ≥ min{l1/2, l2/2}, where l1 and l2 are the length of the coiled-coil segments as annotated in the two ways). Our method correctly predicts the two coiled-coil regions of the HRSV protein fusion core. Furthermore, it is also able to correctly identify the two coiled-coil regions of the SIV protein core, which are not yet reported in the Uniprot database in spite of being present in the protein 3D structure (PDB id: 2ezoA). CCHMM_PROF correctly recognizes one of the two regions of the core structure of the gp2 of Ebola virus. The CCHMM_PROFILE performance on viral proteins is particularly interesting, highlighting its generalizing capabilities. Up to now, only a specialized predictor of viral coiled-coils (LearnCoil-VMF; Singh et al., 1999) could be adopted to predict this kind of coiled-coil segments.

3 DISCUSSION

We describe a new implementation for predicting coiled-coil domains as an approach to select proteins that contain coiled-coil segments in a given set of protein sequences (or in proteomes) and to correctly locate the coiled-coil domains in a protein chain. We build a HMM that takes the evolutionary information obtained from multiple sequence alignments as input. Our results indicate that the introduction of our HMM model that implements sequence profile matrices and machine learning methods. We also provide a new structurally annotated and freely available benchmark dataset of coiled-coil structures that can be used to reliably train and test computational methods. Furthermore, we suggest a more robust evaluation of the method performances, by introducing a new scoring frame that takes into account not only the residue accuracy, but also the accuracy at the segment and at the protein levels. The development of accurate computational methods for coiled-coil prediction can drive experiments towards the de novo design of allo-helical coiled-coil structures and can be applied to genome-wide annotation processes.

Funding: FIBB 2003 LBB1-International Laboratory of Bioinformatics (to R.C.); Bologna node of the Biosapiens Network of Excellence project within the European Union’s VI Framework Programme (contract number LSGH-CT-2003-503265 to R.C.).

Conflict of Interest: none declared.

REFERENCES


