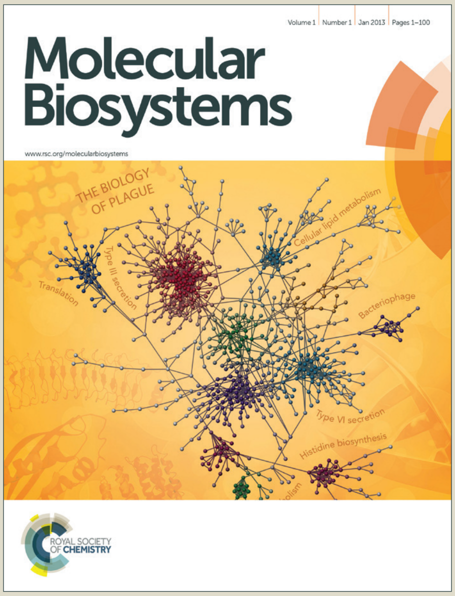
Molecular BioSystems

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular Biosystems

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/molecularbiosystems

Dynamic Article Links ►

Methods

Computational Characterization of Parallel Dimeric and Trimeric Coiled-coils Using Effective Amino Acid Indices †

Chen Li,[a] Xiao-Feng Wang,[b,c] Zhen Chen,[b] Ziding Zhang *[b] and Jiangning Song *[a,d]

Received XX, Accepted XX

DOI:

The coiled-coil, which consists of two or more α-helices winding around each other, is a ubiquitous and most frequently observed protein-protein interaction motif in nature. The coiled-coil is known for its straightforward heptad repeat pattern and can be readily recognized based on protein primary sequences, exhibiting a variety of oligomer states and topologies. Due to the stable interaction formed between their α-helices, coiled-coils have been under close scrutiny to design novel protein structures for potential applications in the fields of material science, synthetic biology and medicine. However, their broader application requires an in-depth and systematic analysis of the sequence-to-structure relationship of coiled-coil folding and oligomeric formation. In this article, we propose a new oligomerization state predictor, termed as *RFCoil*, which exploits the most useful and non-redundant amino acid indices combined with the machine learning algorithm - random forest (RF) to predict the oligomeric states of coiled-coil regions. Benchmarking experiments show that *RFCoil* achieves an AUC (area under the ROC curve) of 0.849 on the 10-fold cross-validation test using the training dataset and 0.855 on the independent test using the validation dataset, respectively. Performance comparison results indicate that *RFCoil* outperforms four existing predictors LOGICOIL, PrOCoil, SCORER 2.0 and Multicoil2. Furthermore, we extract a number of predominant rules from the trained RF model that underlie the oligomeric formation. We also present two case studies to illustrate the applicability of the extracted rules to the prediction of coiled-coil oligomerization state. The *RFCoil* web server, source codes and datasets are freely available for academic users at http://protein.cau.edu.cn/RFCoil/.

Introduction

The coiled-coil is a ubiquitous structural motif consisting of two or more α-helices, which wind around each other to form a rope-like structure. Nearly sixty years ago, Crick proposed the standard structure model of the coiled-coil, which is distinct from other protein structures. Dimeric and trimeric coiled-coils are the two most common types of coiled-coil structures. Coiled-coils can be found in all organisms and it is estimated that nearly 10% and 2-9% of eukaryotic and prokaryotic proteins harbour coiled-coil domain[1-4], respectively. Due to their ability to oligomerize, coiled-coils play crucial roles in many biological processes, such as transcription, intracellular trafficking, viral infection and cellular signaling[5,6]. The property of coiled-coils which enables two proteins to interact with each other, also attracts a great deal of protein designers' interests[7]. Coiled-coils are among the first designed proteins[8,9], with potential applications in material science, synthetic biology and medicine[10,11]. Accordingly, understanding the mechanism of coiled-coil oligomerization is critically important for researchers to design versatile proteins with different functions.

The rope-like structure of coiled coils enables them to generate an interesting heptad repeat sequence pattern. That is, the structure goes around two complete turns of the helix after 7 residues, rather than the regular 7.2 residues. The heptad repeat is often labeled as *abcdefg*. Residues at register positions *a* and *d* are often hydrophobic, forming a buried hydrophobic surface and providing the driving force for oligomerization. In contrast, residues at positions *e* and *g* are often charged or polar, which form salt bridges and electrostatic interactions, helping specify the binding partners[12]. Despite the simple heptad repeat pattern at the sequence level, coiled-coils display a great variety of oligomerization states, including dimers, trimers, tetramers, pentamers, and even heptamers. In addition, they often vary in the helix orientation, parallel or anti-parallel. Most coiled-coils adopt left-handed super-coils; however, right-handed coiled-coils are also observed[13]. Accordingly, an important question to address is, how can this simple heptad sequence repeat pattern encode such diverse structures?

To answer this question, a number of computational methods have been developed to analyze coiled-coils, which can be generally grouped as sequence-based or structure-based methods. Sequence-based methods mainly use the frequencies of residues or residue pairs at specific register positions to predict coiled-coil regions[14-21], oligomerization states[4,17,18,22,23] and helix orientations [24]. In contrast, structure-based methods usually utilise structural information to facilitate the prediction, including SOCKET[12] and Twister[25]. In particular, the SOCKET algorithm is able to recognize characteristic knobs-into-holes side-chain packing of coiled-coil structures, clearly define coiled-coil helix boundaries, oligomerization states and helix orientations and assign heptad registers. The CC+ database[26] is developed based on the SOCKET algorithm, which includes several coiled-coil datasets previously used as training datasets for building coiled-coil classifiers. Twister is implemented to compute local structural parameters of coiled-coils, based on Crick's parameterization[27].

Regarding the prediction of coiled-coil oligomerization state, two early-stage algorithms SCORER[28] and Multicoil[29] exist. More recently, two new versions, SCORER 2.0[23] and Multicoil2[17] have been developed, and shown to perform better than their respective older versions. Almost at the same time, another two predictors for coiled-coil oligomerization state, PrOCoil[22] and LOGICOIL[4], were published. Multicoil2 employs a Markov Random Field method to integrate sequence features. It assigns the probability of a residue in a sequence to be non-coiled-coil, dimeric or trimeric. SCORER 2.0 and PrOCoil classify parallel dimeric and trimeric coiled-coils, given a coiled-coil sequence with known heptad registers. SCORER2.0 uses statistically significant amino acid frequencies at seven heptad registers in combination with a Bayes factor method to distinguish parallel dimers from trimers. PrOCoil designs a new kernel function and uses the SVM (Support Vector Machine) algorithm to classify parallel dimmers and trimers[22]. LOGICOIL, trained with coiled-coil regions larger than 14 amino acids by Bayesian variable selection response probabilities, can predict multiple oligomerization states for coiled-coil regions such as parallel dimer, antiparallel dimer, trimer and tetramer[4]. Therefore, LOGICOIL is currently considered as the state-of-the-art predictor for oligomerization states of coiled-coils.

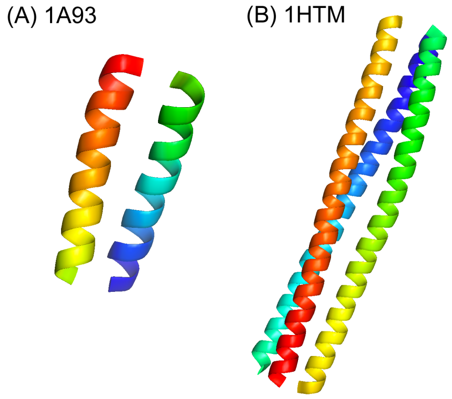


Fig. 1 Cartoon representations of parallel (A) dimeric (PDB ID: 1A93[30]) and (B) trimeric (PDB ID: 1HTM[31]) coiled-coils.

In this article, we address the same classification task of SCORER 2.0 and PrOCoil by developing a novel tool *RFCoil*, which uses a sequence-based approach to distinguish parallel dimeric from trimeric coiled-coils (See Fig. 1 for examples of parallel dimer and trimer). More specifically, *RFCoil* employs the random forest (RF) algorithm to identify the most important and non-redundant amino acid indices and construct the classifiers to predict the oligomerization state of coiled-coils. We further compare the performance of *RFCoil* with four existing tools SCORER 2.0, PrOCoil, Multicoil2 and LOGICOIL by performing both 10-fold cross-validation and independent tests. The results show that *RFCoil* outperforms four existing tools LOGICOIL, SCORER 2.0, PrOCoil and Multicoil2 on the independent test. Moreover, we extract a number of important rules from the built RF models in an effort to provide biological insights into the underlying rules of the formation of oligomerization states of coiled-coils.

Materials and Methods

Dataset

We used the benchmark dataset originally compiled by the developers of PrOCoil to train our models and assess the performance of our method. This benchmark dataset comprises 385 dimers and 92 trimers. The minimum length of the coiled-coils is 8 and nearly half of the coiled-coils have lengths longer than 14. This dataset was further divided into ten folds, and any two sequences from different folds have a sequence identity no more than 60%. The methods were tested using the 10-fold cross-validation tests.

Moreover, apart from the benchmark dataset, we also constructed an independent test dataset to assess and compare the predictive performance of different methods. The procedures for constructing this independent test dataset are as follows: First, we used the SOCKET algorithm[12] to search the PDB database[32] for parallel coiled-coil dimers and trimers. For dimers, we selected those sharing a sequence identity of no more than 60% with the dimeric coiled-coil sequences in the training dataset. The selected dimers were further filtered to ensure that any two sequences shared a sequence identity of no more than 60%. The trimeric coiled-coils were filtered in a similar way as the dimers. Note that the sequence identity was calculated using the Needleman-Wunsch algorithm[33]. The final independent test set consists of 363 dimers and 48 trimers.

RFCoil

Our *RFCoil* approach includes four major steps, as shown in Fig. 2. The first step is to construct the training and independent test datsets extracted from the PDB database. The second step is to encode the input data, which was achieved by extracting the average amino acid index values for each heptad register. The third step is to select the informative and non-redundant features for oligomerization state classification. We assumed that no prior knowledge of each feature's importance was known and this makes it possible that our feature selection method presented here can be applied to other questions. The final step is to use the selected features as the input to train *RFCoil* models. More details about the *RFCoil* approach are discussed in the following sections.

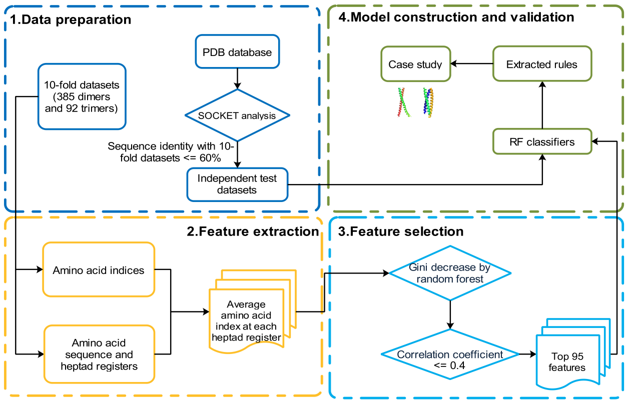

1.Data preparation
PDB database
10-fold datasets (385 dimers and 92 trimers)
SOCKET analysis
Sequence identity with 10-fold datasets <= 60%
Independent test datasets
4.Model construction and validation
Case study
Extracted rules
RF classifiers
2.Feature extraction
Amino acid indices
Amino acid sequence and heptad registers
Average amino acid index at each heptad register
3.Feature selection
Gini decrease by random forest
Correlation coefficient <= 0.4
Top 95 features

Fig. 2 Flowchart of *RFCoil*. Its development comprises four major steps, including data preparation, feature extraction, feature selection and RF model training and validation.

Sequence encoding

We attempted to capture the coiled-coil's oligomerization state using its amino acid sequence information and each coiled-coil sequence using the physiochemical and biochemical properties of amino acids. To realize this, we extracted 529 amino acid indices that had no "NA" values in the AAindex database[34] (see Table S1 and S2, ESI†). We encoded each coiled-coil sequence using the average amino acid index value at each heptad register, obtained by the following equation:

$$I(r,i) = \frac{\sum_{a\in r} AA(a,i)}{n(r)} \qquad (1)$$

where r represents a heptad register which can be $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{e}$, $\boldsymbol{f}$ or $\boldsymbol{g}$, i denotes the *i-th* amino acid index amongst the 529 amino acid indices, a represents the amino acid residue in the coiled-coil sequence whose heptad register is r, $AA(a, i)$ stands for the value of the *i-th* amino acid index for the amino acid a, while $n(r)$ is the number of amino acid residues at the heptad register r. As there are a total of 7 heptad registers and 529 amino acid indices, a coiled-coil sequence is represented by a 3703-dimensional vector.

Random forest

Ensemble learning is a prevalent machine learning technique. Its underlying principle is based on the observation that the ensemble of some weak classifiers can usually achieve a better accuracy than a single classifier when using the same training information. RF[35] is an effective ensemble learning algorithm and has been widely applied in bioinformatics[36-41]. RF consists of many decision trees, each of which is grown as follows. Suppose that there are N instances and M variables in the training set. First, N instances are randomly selected from the training set with replacement. Second, at each node, $\sqrt{M}$ variables are randomly selected and the best is used to split the node. Finally, each tree is grown as large as possible. The RF chooses the classification of the most votes given by all the individual trees. In this work, the random forest algorithm was implemented using the 'randomForest' R package[42].

Feature selection and model training

As described above, a coiled-coil sequence was encoded by 3703 features. However, it is likely that some features are irrelevant or redundant, making little or no contribution to the prediction. We thus performed feature selection experiments to select and identify the most meaningful features for the classification of coiled-coil oligomerization states. For each feature, i.e. the variable in the RF, its importance is measured by the gini index of RF. When splitting the variable on a node in the process of growing a tree, the gini impurity criterion, which is a "goodness of split" criterion[43], is less than the parent node for the two child nodes. Therefore, summing up the gini decrease for the variables over all trees gives the value to assess the variable's importance.

After evaluating each feature's importance, another issue remains to be resolved. That is, the integration of individual best features does not necessarily lead to the best classification performance[44] and there still exists redundancy between different features. For example, there are many amino acid indices that describe the amino acid hydrophobicity in the AAindex database and some might be highly correlated with each other. To address this, we calculated the correlation coefficient between any two amino acid indices. If two features encode the same heptad register and the correlation coefficient of their representative amino acid indices has an absolute value of less than a threshold c, then the feature with a smaller gini decrease will be removed from the feature set. After this repetitive procedure, we select the top n features to build the final RF model.

In the above process, we used the Kendall rank correlation coefficient. Let $(X_1, X_2, \ldots, X_{20})$ and $(Y_1, Y_2, \ldots, Y_{20})$ be two sets of amino acid indices. A pair of amino acid index values (X_i, Y_i) and (X_j, Y_j) are defined to be concordant, if both $X_i > X_j$ and $Y_i > Y_j$ or both $X_i < X_j$ and $Y_i < Y_j$, or defined to be discordant, if $X_i > X_j$ and $Y_i < Y_j$ or $X_i < X_j$ and $Y_i > Y_j$. The Kendall correlation coefficient τ is defined as follows:

$$\tau = \frac{nc-nd}{\frac{1}{2}\times 20\times(20-1)} \qquad (2)$$

where nc and nd represent the numbers of concordant pairs and discordant pairs, respectively.

Extracting signifcant rules

Each tree in the RF can be represented by a set of rules. Each path from the root to a leaf node in a tree is a rule. A totoal of 4,000 decision trees were grown in our work to build the RF model, resulting in many rules present in the model. We devised a method to extract a rule set that contains as few rules as possible to correctly classify all the instances in the dataset: Firstly, we extracted the rules without wrongly classifying any instance in the dataset and identified the rules that could classify the largest number of dimers or trimers; Secondly, we saved the rules found in the first step in the rule set and removed those instances that were correctly classified by the rule; Thirdly, we repeated steps 1 and 2 until there were no instances in the dataset.

Accessing the prediction performance of RF model

We used the receiver operating characteristic (ROC) curve[45] to assess the prediction performance of RF model. The ROC curve is a plot of true positive rate (TPR) against false positive rate (FPR). TPR defines the ratio of correctly predicted positives to all the positive instances, while FPR stands for the ratio of incorrectly predicted positives to all the negative instances. In this study, we defined dimeric coiled-coils as positive instances and trimeric coiled-coils as negative instances. In addition, the area under the ROC curve (AUC) represents the probability for a classifier to rank a randomly selected positive instance higher than a randomly selected negative one. Hence, AUC was also used as an important performance measure in this study to compare the performance of different methods.

Performance comparison between *RFCoil* and four existing predictors

To evaluate the performance of *RFCoil*, we conducted two benchmarking experiments. In the first benchmarking experiment, we compare the performance of *RFCoil* with SCORER 2.0 and PrOCoil by performing 10-fold cross-validation tests on the PrOCoil dataset. In the second benchmarking experiment, we used the PrOCoil dataset as the training dataset to train the models of *RFCoil* and PrOCoil. Then the constructed independent test dataset was used to assess the performance of

RFCoil in comparsion with the other four tools SCORER 2.0, PrOCoil, Multicoil2 and LOGICOIL. In particular, the prediction outputs of SCORER 2.0, PrOCoil and LOGICOIL were generated by their local versions downloaded from the corresponding websites. In the case of Multicoil2, we instead submitted the test sequences to its online server and obtained the prediction results.

Results and discussion

In this section, we first report the prediction performance of *RFCoil* in comparison to SCORER 2.0 and PrOCoil on the 10-fold cross-validation tests. We then comprehensively assess the performance of *RFCoil*, PrOCoil, SCORER 2.0, LOGICOIL and Multicoil2 on the independent tests. Finally, we discuss the final features selected by our feature selection method and the extracted significant rules on the PrOCoil benchmark dataset.

Prediction performance on the 10-fold cross-validation tests using the PrOCoil dataset

We performed 10-fold cross-validation tests to assess the performance of the predictive models of *RFCoil* using the PrOCoil dataset (Table 1). When using the average amino acid index values at each heptad as the input, the average AUC of *RFCoil* was 0.819, compared with 0.808 of PrOCoil and 0.789 of SCORER 2.0, respectively. After setting the Kendall correlation coefficient between the amino acid indices at ≤ 0.4 to select the 95 top features, the average AUC of *RFCoil* was further improved to 0.849. The authors of PrOCoil[22] found that the training set could be further augmented by blast search against the NCBI-NR database, which could provide an improved prediction performance in their study. Here, our results indicate that the AUC of PrOCoil on the augmented training dataset indeed reached 0.818, representing a better performance than that of the original PrOCoil. On the other hand, we find that *RFCoil* performed the best for certain folds and reasonably well for other folds during 10-fold cross-validation tests (Table 1). In summary, *RFCoil* achieved a better performance than the other two methods PrOCoil and SCORER 2.0 on the 10-fold cross-validation tests using the PrOCoil dataset. According to the 10-fold cross-validation tests, we implemented the final online web server of *RFCoil* using the selected feature set.

Table 1 The AUC scores of *RFCoil*, SCORER2.0 and PrOCoil, evaluated using 10-fold cross-validation tests.

Fold	*RFCoil* (all features)	*RFCoil* (selected features)	SCORER2.0	PrOCoil	PrOCoil_blast[a]
1	0.612	0.691	0.773	0.882	0.882
2	0.801	0.817	0.776	0.967	0.935
3	0.750	0.835	0.625	0.581	0.681
4	0.885	0.875	0.810	0.830	0.850
5	0.971	0.957	0.833	0.848	0.867
6	0.869	0.865	0.808	0.741	0.842
7	0.908	0.961	0.875	0.809	0.724
8	0.803	0.769	0.735	0.744	0.744
9	0.698	0.825	0.651	0.738	0.702
10	0.890	0.895	1.000	0.943	0.957
Average	0.819	0.849	0.789	0.808	0.818

[a] PrOCoil_blast denotes the model trained using the augmented PrOCoil dataset by blast search against NCBI-NR database.

Table 2 Statistics of the selected features

Heptad register	***a***	***b***	***c***	***d***	***e***	***f***	***g***
Number of features	13	5	8	10	9	5	8
Sum of the gini decrease	35.6	5.8	12.5	16.6	18.4	7.8	11.8

Prediction performance on the independent tests

In addition to the performance evaluation using the PrOCoil benchmark dataset, we also curated an independent test dataset to comprehensively compare the performance of our method *RFCoil* for predicting coiled-coil oligomerization state with four existing predictors SCORER 2.0, PrOCoil, Multicoil2 and LOGICOIL. In particular, we used the PrOCoil dataset as the training set to build the two types of predictive models of *RFCoil* (denoted as "*RFCoil* (all features)" and "*RFCoil* (selected features)" which used all features and final selected features as the respective inputs to build the models) to classify coiled-coil sequences in this independent test dataset. LOGICOIL and SCORER 2.0 were trained on the coiled-coil sequences no shorter than 15 amino acids, while Multicoil2 could only predict coiled-coil sequences longer than 21 amino acids. In the case of PrOCoil, it requires a minimum length of coiled-coil sequences of 8 amino acids. In this study, we reported the results by performing the independent test using the PrOCoil dataset.

The output scores were selected from two prediction categories of LOGICOIL (i.e., parallel dimer and trimer) and normalized to [0,1] before plotting the ROC curve. Instead of providing an overall prediction score for the input sequence, Multicoil2 provides predicted probabilities for each individual residue in the sequence of forming dimers, trimers or non-coiled-coils. Accordingly, to compare with other methods, we calculated the average of the predicted probabilities of Multicoil2, normalized them into the range of [0,1] and removed the predicted non-coiled-coils from the results (with the prediction threshold set at 0.5).

The ROC curves and the corresponding AUC values of *RFCoil*, SCORER 2.0, PrOCoil, LOGICOIL and Multicoil2 on the independent tests are shown in Fig. 3. The AUC values of the two types of *RFCoil* models that used all features and the final selected features as inputs were 0.855 and 0.851, respectively. These represent the overall best AUC scores among different predictors. In contrast, Multicoil2 achieved an AUC value of 0.689, while SCORER 2.0 achieved an AUC score of 0.776. PrOCoil achieved an AUC value of 0.736 and the PrOCoil_blast model trained using the augmented dataset achieved an AUC of 0.723, both of which decreased considerably compared to that on the 10-fold cross validation. In contrast, LOGICOIL achieved an AUC value of 0.757. We also noted that augmenting the training set in this case did not help improve the performance of PrOCoil, as reflected by a lower AUC of 0.723 by the latter model.

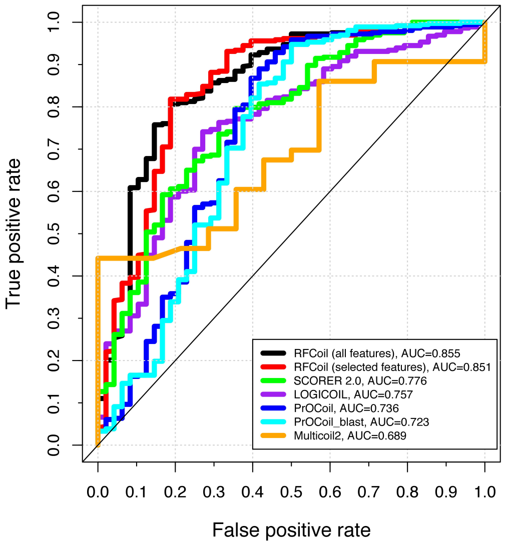


Fig. 3 The ROC curves of different methods on independent test dataset.

Analysis of final selected features based on the PrOCoil dataset

Application of the Kendall correlation coefficient set at ≤ 0.4 resulted in a subset of top 95 features selected (see Table S3, ESI†). The average AUC of the *RFCoil* model trained using this selected feature set reached its maximum value of 0.849 on the 10-fold cross validation tests using the PrOCoil benchmark dataset (Table 1). We further calculated the number of features at each heptad register, as well as the sum of the gini decreases for the features at each heptad register. Table 2 shows that the position *a* is the most important position for the discrimination between parallel dimers and trimmers, as determined by the sum of the gini decreases. The other positions *d*, *e*, *c*, *g* are less important compared with the position *a*, while positions *f* and *b* are the least important positions.

Significant rules extracted from the PrOCoil dataset

Using the method of rule extraction described in the Methods section, we extracted 10 significant rules covering all the 382 dimers, and another 10 significant rules covering all the 92 trimers in the PrOCoil dataset. The description of each specific rule and the numbers of dimers and trimers covered by the corresponding rule are given in Tables 3 and 4, respectively. Note that it is likely that a sample in the dataset may be identified by more than two rules, as shown in the tables.

Each rule is a combination of useful amino acid indices at certain heptad registers. The RF algorithm is particularly powerful in making use of the correlations between different heptad registers for efficient classification. In contrast, SCORER2.0 only uses residue frequencies at each heptad register, failing to take into account the potential interactions between different heptad-repeat positions, while PrOCoil employs the frequencies of each amino acid pair in each pair of heptad registers. An important advantage of RF is that it can make use of the correlations between two or more heptad registers. This might explain why our method outperformed the other four methods PrOCoil, Multicoil 2, SCORER 2.0 and LOGICOIL.

Case studies

Using the selected 95 features on the PrOCoil dataset, we built the RF model and illustrated the performance of this model on two parallel coiled-coil structures from the independent test dataset (see Fig. S1 for structural information of these two proteins, ESI†). The first one is a coiled-coil parallel dimer from the Rho-associated protein kinase 1 (PDB ID: 3O0Z). This protein is involved in a variety of cellular processes including muscle contraction, cell migration and stress fiber formation[46]. Its predicted probability for being dimeric by the RF model was 0.872. The other is a trimer from the avian reovirus S1133 fibre (PDB ID: 2VRS), a minor component of the avian reovirus outer capsid[47]. Its probability for being parallel trimer predicted by the RF model was 0.759. The coiled-coil oligomerization states of both proteins were correctly predicted by *RFCoil*.

In addition, we found that the dimeric coiled-coil in the Rho-associated protein kinase 1 conformed to the significant rules 1, 2, 5 and 10, as listed in Table 3. Further, the trimeric coiled-coil in 2VRS conformed to the significant rules 1 and 5 listed in Table 4. Altogether, these results showcase the predictive ability of the constructed *RFCoil* model and usefulness of the extracted rules based on the selected effective amino acid indices.

Conclusions

In this article, we addressed the challenging task of distinguishing parallel dimeric from trimeric coiled-coils by developing an RF-based approach termed as *RFCoil*, which used effective amino acid indices to build the predictive models. To remove redundant and irrelevant features and improve the classification performance, we combined the gini index calculated by RF and the correlation coefficients between the amino acid indices at different positions of heptad registers to select the most meaningful features. The model trained using the selected features indeed improved the prediction performance. We further analyzed the selected features and proposed a rule extraction method to identify significant rules from the RF model to better understand important rules that underlie the organization of dimeric and trimeric coiled-coils. The rules provide useful insights into the design of coiled-coil proteins. In addition, our method can be readily extended to predict coiled-coils of higher order oligomerization states, provided that more solved structures are available in the near future. Benchmarking experiments indicate that *RFCoil* outperforms the other four exisitng tools. It is expected to become an efficient tool to facilitate the studies of coiled-coil structures. Finally, as an implementation of our method, an online prediction server of *RFCoil* is made freely available at http://protein.cau.edu.cn/RFCoil. The source code can be downloaded for interested users to build their specific models using their own datasets.

Table 3 The extracted rules for coiled-coil dimers

No.	Description of the rule[a]	Number of samples covered by the rule
1	I(***c***, 260)<=0.2825 & I(***d***, 17)>4.2435 & I(***f***, 16)>7.213 & I(***f***, 240)>-3.3475 & I(***a***, 294)>-0.2925 & I(***a***, 400)<=14.183	225
2	I(***c***, 340)<=5.83 & I(***d***, 17)>4.2315 & I(***d***, 195)>2.1225 & I(***e***, 74)>-62.35 & I(***f***, 16)<=8.676 & I(***f***, 73)>240.0835 & I(***g***, 50)<=0.088 & I(***g***, 201)<=1.654 & I(***g***, 408)>1.1735 & I(***b***, 18)<=7.3085 & I(***b***, 273)>-0.375	173
3	I(***c***, 371)>0.355 & I(***d***, 220)<=2.9275 & I(***e***, 372)<=2.202 & I(***e***, 495)<=0.9795 & I(***g***, 61)>0.3625 & I(***a***, 386)<=0.388	128
4	I(***e***, 299)<=1.165 & I(***a***, 275)>0.1125 & I(***a***, 374)<=0.7665	58
5	I(***c***, 194)>-1.4475 & I(***c***, 293)<=0.4225 & I(***c***, 340)<=4.169 & I(***d***, 342)>-1.2415 & I(***d***, 401)<=1.22 & I(***f***, 338)<=1.4625 & I(***g***, 155)>107.1895 & I(***g***, 201)>0.759 & I(***a***, 44)>0.5575 & I(***b***, 529)>-3.1775	201
6	I(***c***, 303)<=1.2345 & I(***d***, 17)>4.279 & I(***d***, 342)>-0.425 & I(***e***, 110)>0.3625 & I(***g***, 336)<=0.8415 & I(***a***, 386)<=0.1705 & I(***a***, 506)>1.4695	34
7	I(***c***, 18)<=6.9585 & I(***c***, 361)>-0.177 & I(***e***, 296)<=0.2385 & I(***a***, 400)<=16.35 & I(***a***, 506)<=1.7425 & I(***b***, 185)<=4.195 & I(***a***, 99)<=1.54	183
8	I(***a***, 107)>0.7325 & I(***d***, 401)<=1.21 & I(***a***, 1)<=4.7025 & I(***a***, 294)>-0.335 & I(***g***, 370)<=0.773 & I(***b***, 185)<=4.195	185
9	I(***c***, 326)<=1.5165 & I(***e***, 296)<=0.28 & I(***e***, 495)<=0.9985 & I(***g***, 408)<=1.171	60
10	I(***c***, 18)>6.89 & I(***c***, 141)>0.45 & I(***d***, 94)>0.8835 & I(***d***, 275)<=0.097 & I(***e***, 296)>0.161 & I(***f***, 331)<=1.2875 & I(***g***, 61)<=1.056 & I(***a***, 337)>0.7415	9

[a] "&" denotes the conjunction word "and", while *I(r, n)* represents the *n-th* amino acid index at the heptad *r*.

Table 4 The extracted rules of coiled-coil trimers

No.	Description of the rule[a]	Number of samples covered by the rule
1	I(***c***, 236)>0.795 & I(***c***, 361)<=0.123 & I(***d***, 326)<=0.7415 & I(***e***, 219)>0.945 & I(***e***, 299)>1.1665 & I(***g***, 201)>0.536 & I(***g***, 309)>0.8665 & I(***a***, 400)>14.1515 & I(***a***, 506)>1.464	44
2	I(***c***, 293)>-0.324 & I(***c***, 361)<=0.123 & I(***c***, 405)<=1.2725 & I(***d***, 175)<=0.8575 & I(***e***, 110)>0.3725 & I(***f***, 16)<=8.5555 & I(***g***, 408)>0.655 & I(***a***, 374)<=0.826 & I(***b***, 529)<=-3.167 & I(***b***, 273)>-0.1685	43
3	I(***c***, 340)>0.096 & I(***a***, 176)>0.675 & I(***d***, 195)>5.3525 & I(**f**, 73)<=245.6 & I(***f***, 385)>-0.0975 & I(***a***, 386)<=0.1365 & I(***b***, 18)>5.85	17
4	I(***d***, 94)<=1.154 & I(***a***, 18)>5.125 & I(***g***, 336)>0.8415 & I(***a***, 374)<=0.765 & I(***a***, 400)>12.6765 & I(***b***, 18)<=7.7415 & I(***b***, 273)>-0.104	30
5	I(***c***, 361)<=-0.176 & I(***d***, 94)<=1.2915 & I(***a***, 176)<=0.8375 & I(***e***, 360)<=0.2115 & I(***b***, 329)<=1.325	19
6	I(***c***, 141)>0.655 & I(***e***, 296)>0.2665 & I(***g***, 408)<=0.9935 & I(***a***, 374)>0.7135	8
7	I(***a***, 107)>0.7505 & I(***d***, 220)<=2.9275 & I(***d***, 240)<=-2.141 & I(***e***, 495)>0.9795 & I(***a***, 294)<=-0.245	5
8	I(***d***, 195)<=9.1325 & I(***d***, 422)>-0.501 & I(***e***, 295)>-0.061 & I(***e***, 372)>0.1965 & I(***f***, 73)<=267.9165 & I(***a***, 374)>0.6285 & I(***a***, 400)>14.385 & I(***a***, 99)<=1.2675	16
9	I(***c***, 340)>-0.0625 & I(***c***, 371)>1.061 & I(***f***, 16)>8.481 & I(***g***, 12)>-4.7165 & I(***b***, 284)>-0.06	14
10	I(***b***, 478)>1.6165 & I(***c***, 361)>-0.1935 & I(***d***, 74)>-25.1175 & I(***d***, 422)>-0.3215 & I(***g***, 98)>1.0125 & I(***g***, 370)<=0.773	3

[a] See the footnote in Table 3 for the notations of each symbols in the rules.

Acknowledgments

This work is supported by Hundred Talents Program of the Chinese Academy of Sciences (CAS), National Natural Science Foundation of China (No. 61202167, 61303169, 11250110508, 31350110507), Knowledge Innovation Program of CAS (No. KSCX2-EW-G-8), Tianjin Municipal Science & Technology Commission (No. 10ZCKFSY05600), Major Inter-Disciplinary Research (IDR) Project awarded by Monash University and National Health and Medical Research Council of Australia (NHMRC) (No. 490989). JS is an NHMRC Peter Doherty Fellow and recipient of the Hundred Talents Program of CAS.

Notes and references

[a] *Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia*
[b] *State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China*
[c] *School of Mathematics and Computer Science, Shanxi Normal University, Linfen 041004, China*
[d] *National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China*

† Electronic supplementary information (ESI) available
* Corresponding authors.
Email: ZZ (zidingzhang@cau.edu.cn); JS (jiangnng.song@monash.edu)

1. G. Grigoryan and A. E. Keating, *Curr. Opin. Struct. Biol.*, 2008, **18**, 477-483.
2. T. L. Vincent, D. N. Woolfson and J. C. Adams, *Int. J. Biochem. Cell Biol.*, 2013, **45**, 2392-2401.
3. A. A. McFarlane, G. L. Orriss and J. Stetefeld, *Eur. J. Pharmacol*, 2009, **625**, 101-107.
4. T. L. Vincent, P. J. Green and D. N. Woolfson, *Bioinformatics*, 2013, **29**, 69-76.
5. A. N. Lupas and M. Gruber, *Adv. Protein Chem.*, 2005, **70**, 37-78.
6. Y. Wang, X. Zhang, H. Zhang, Y. Lu, H. Huang, X. Dong, J. Chen, J. Dong, X. Yang, H. Hang and T. Jiang, *Mol. Biol. Cell*, 2012, **23**, 3911-3922.
7. N. R. Zaccai, B. Chi, A. R. Thomson, A. L. Boyle, G. J. Bartlett, M. Bruning, N. Linden, R. B. Sessions, P. J. Booth, R. L. Brady and D. N. Woolfson, *Nat. Chem. Biol.*, 2011, **7**, 935-941.
8. S. F. Betz, J. W. Bryson and W. F. DeGrado, *Curr. Opin. Struct. Biol*, 1995, **5**, 457-463.
9. K. Chen and L. Kurgan, *Methods Mol. Biol.*, 2013, **932**, 63-86.
10. A. Lupas, *Trends Biochem. Sci.*, 1996, **21**, 375-382.
11. E. H. Bromley, K. Channon, E. Moutevelis and D. N. Woolfson, *ACS Chem. Biol.*, 2008, **3**, 38-50.
12. J. Walshaw and D. N. Woolfson, *J. Mol. Biol.*, 2001, **307**, 1427-1450.
13. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber and P. S. Kim, *Science*, 1998, **282**, 1462-1467.
14. L. Bartoli, P. Fariselli, A. Krogh and R. Casadio, *Bioinformatics*, 2009, **25**, 2757-2763.
15. O. J. Rackham, M. Madera, C. T. Armstrong, T. L. Vincent, D. N. Woolfson and J. Gough, *J. Mol. Biol.*, 2010, **403**, 480-493.
16. M. Delorenzi and T. Speed, *Bioinformatics*, 2002, **18**, 617-625.
17. J. Trigg, K. Gutwin, A. E. Keating and B. Berger, *PLoS One*, 2011, **6**, e23519.
18. E. Wolf, P. S. Kim and B. Berger, *Protein Sci.*, 1997, **6**, 1179-1189.
19. A. V. McDonnell, T. Jiang, A. E. Keating and B. Berger, *Bioinformatics*, 2006, **22**, 356-358.
20. B. Berger, D. B. Wilson, E. Wolf, T. Tonchev, M. Milla and P. S. Kim, *Proc. Natl. Acad. Sci. U S A*, 1995, **92**, 8259-8263.
21. A. Lupas, M. Van Dyke and J. Stock, *Science*, 1991, **252**, 1162-1164.
22. C. C. Mahrenholz, I. G. Abfalter, U. Bodenhofer, R. Volkmer and S. Hochreiter, *Mol. Cell. Proteomics*, 2011, **10**, M110 004994.
23. C. T. Armstrong, T. L. Vincent, P. J. Green and D. N. Woolfson, *Bioinformatics*, 2011, **27**, 1908-1914.
24. J. R. Apgar, K. N. Gutwin and A. E. Keating, *Proteins*, 2008, **72**, 1048-1065.
25. S. V. Strelkov and P. Burkhard, *J. Struct. Biol.*, 2002, **137**, 54-64.
26. O. D. Testa, E. Moutevelis and D. N. Woolfson, *Nucleic Acids Res.*, 2009, **37**, D315-322.
27. F. H. Crick, *Acta Crystallogr.*, 1953, **6**, 689-697.
28. D. N. Woolfson and T. Alber, *Protein Sci.*, 1995, **4**, 1596-1607.
29. P. S. Kim, B. Berger and E. Wolf, *Protein Sci.*, 1997, **6**, 1179-1189.
30. P. Lavigne, M. P. Crump, S. M. Gagne, R. S. Hodges, C. M. Kay and B. D. Sykes, *J. Mol. Biol.*, 1998, **281**, 165-181.
31. P. A. Bullough, F. M. Hughson, J. J. Skehel and D. C. Wiley, *Nature*, 1994, **371**, 37-43.
32. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235-242.
33. S. B. Needleman and C. D. Wunsch, *J Mol. Biol.*, 1970, **48**, 443-453.
34. S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama and M. Kanehisa, *Nucleic Acids Res.*, 2008, **36**, D202-205.
35. L. Breiman, *Mach. Learn.*, 2001, **45**, 5-32.
36. X. F. Wang, Z. Chen, C. Wang, R. X. Yan, Z. Zhang and J. Song, *PLoS One*, 2011, **6**, e26767.
37. C. Zheng, M. Wang, K. Takemoto, T. Akutsu, Z. Zhang and J. Song, *PLoS One*, 2012, **7**, e49716.
38. M. M. Dehmer, N. N. Barbarini, K. K. Varmuza and A. A. Graber, *BMC Struct. Biol.*, 2010, **10**, 18.
39. S. Hirose, K. Yokota, Y. Kuroda, H. Wako, S. Endo, S. Kanai and T. Noguchi, *BMC Struct. Biol.*, 2010, **10**, 20.
40. M. Wang, X. M. Zhao, K. Takemoto, H. Xu, Y. Li, T. Akutsu and J. Song, *PLoS One*, 2012, **7**, e43847.
41. Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang and L. Chen, *Bioinformatics*, 2010, **26**, 1616-1622.
42. A. Liaw and M. Wiener, *R News*, 2002, 2, 18-22.
43. L. Raileanu and K. Stoffel, *Ann. Math. Artif. Intel.*, 2004, **41**, 77-93.
44. H. Peng, F. Long and C. Ding, *IEEE Trans. Pattern. Anal. Mach. Intell.*, 2005, **27**, 1226-1238.
45. T. Fawcett, *Pattern Recogn. Lett.*, 2006, **27**, 861-874.
46. D. Tu, Y. Li, H. K. Song, A. V. Toms, C. J. Gould, S. B. Ficarro, J. A. Marto, B. L. Goode and M. J. Eck, *PLoS One*, 2011, **6**, e18080.
47. P. Guardado-Calvo, G. C. Fox, A. L. Llamas-Saiz and M. J. van Raaij, *J. Gen. Virol.*, 2009, **90**, 672-677.