Bioinformatics, 35(16), 2019, 2790–2795
doi: 10.1093/bioinformatics/bty1062
Advance Access Publication Date: 2 January 2019
Original Paper

Sequence analysis

DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences

Jan Ludwiczak[1,2], Aleksander Winski[1], Krzysztof Szczepaniak[1], Vikram Alva[3] and Stanislaw Dunin-Horkawicz[1,*]

[1]Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland, [2]Laboratory of Bioinformatics, Nencki Institute of Experimental Biology, Warsaw 02-093, Poland and [3]Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen 72076, Germany

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 29, 2018; revised on December 19, 2018; editorial decision on December 21, 2018; accepted on December 27, 2018

Abstract

Motivation: Coiled coils are protein structural domains that mediate a plethora of biological interactions, and thus their reliable annotation is crucial for studies of protein structure and function.
Results: Here, we report DeepCoil, a new neural network-based tool for the detection of coiled-coil domains in protein sequences. In our benchmarks, DeepCoil significantly outperformed current state-of-the-art tools, such as PCOILS and Marcoil, both in the prediction of canonical and non-canonical coiled coils. Furthermore, in a scan of the human genome with DeepCoil, we detected many coiled-coil domains that remained undetected by other methods. This higher sensitivity of DeepCoil should make it a method of choice for accurate genome-wide detection of coiled-coil domains.
Availability and implementation: DeepCoil is written in Python and utilizes the Keras machine learning library. A web server is freely available at https://toolkit.tuebingen.mpg.de/#/tools/deepcoil and a standalone version can be downloaded at https://github.com/labstructbioinf/DeepCoil.
Contact: s.dunin-horkawicz@cent.uw.edu.pl
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Canonical coiled-coil domains consist of two or more α-helices in a parallel or antiparallel orientation that are wrapped around each other into regular left-handed supercoiled bundles. Coiled-coil regions are present in ∼10% of all proteins and support a wide range of biological functions, such as transport of molecules, providing structural rigidity and transduction of conformational changes (Lupas and Bassler, 2017). Owing to their regularity and stability, coiled coils are also of great interest as templates for designing new structures (Woolfson, 2017) and systems for drug delivery (McFarlane *et al.*, 2009), respectively.

The stability of coiled coils stems from the regular meshing of side chains into the so-called knobs-into-holes packing, in which a residue from one helix (knob) packs into a cavity formed by side-chains of the facing helix (hole) (Lupas *et al.*, 2017). Canonical coiled-coil structures, arranged according to the knobs-into-holes packing, are underpinned by seven-residue sequence repeats, referred to as heptad repeats. If the seven positions of a heptad repeat are labeled *a*–*g*, the residues forming the core are in positions *a* and *d*. The core-forming positions are usually occupied by hydrophobic residues, whereas the remaining, solvent-exposed positions are dominated by hydrophilic residues. Repeats longer than seven residues are the basis for the formation of non-canonical coiled coils (Lupas *et al.*, 2017). All such repeats can be described as combinations of three- and four-residue segments. For example, the combination of 3 + 4 + 4 segments leads to a 11-residue repeat (hendecad), characteristic of slightly right-handed coiled-coil bundles. In many natural coiled coils, transitions between different repeats types can be seen; for instance, coiled-coil stalk of the trimeric autotransporter adhesin YadA contains three consecutive repeat types:

non-canonical 15-residue (pentadecad) and 19-residue (nonadecad) repeats followed by a canonical 7-residue repeat (Alvarez *et al.*, 2010). Regardless of the repeat type and its context, the deviation from the canonical heptad repeat causes the appearance of an additional form of interhelical packing termed knobs-into-knobs. The structural constraints imposed by this packing are different than the ones imposed by knob-into-holes packing, but they also lead to the periodical appearance of positions dominated by hydrophobic and hydrophilic residues.

The regular nature of coiled coils makes them a perfect model for studying sequence-structure relationships (Grigoryan and Degrado, 2011; Szczepaniak *et al.*, 2018) and opens the possibility of developing computational methods for the prediction of their structural features based on sequence information. These methods fall into two main categories, namely coiled-coil oligomeric state prediction and coiled-coil domain detection. While some oligomeric state prediction methods such as SCORER (Armstrong *et al.*, 2011), Multicoil2 (Trigg *et al.*, 2011) and RFCoil (Li *et al.*, 2015) allow discrimination between dimers and trimers, others such as LOGICOIL (Vincent *et al.*, 2013) can also predict tetramers and topology of dimers (parallel versus antiparallel). The second category of predictions, i.e. the detection of coiled-coil domains, can be performed based on a single sequence or a sequence profile derived from a multiple sequence alignment. Unsurprisingly, methods that take advantage of evolutionary information (profile-based methods) perform better (for a comparison of the available methods see Gruber *et al.*, 2006; Li *et al.*, 2016).

Here, we describe DeepCoil, a new neural network-based method for the prediction of canonical and non-canonical coiled coils based on a sequence (DeepCoil_SEQ) or a sequence profile (DeepCoil_PSSM). By performing a rigorous benchmark using two independent test sets, we show that both versions of DeepCoil outperform current state-of-the-art methods such as COILS (Lupas *et al.*, 1991), PCOILS (Gruber *et al.*, 2005), Marcoil (Delorenzi and Speed, 2002), Multicoil2 (Trigg *et al.*, 2011) and CCHMM_PROF (Bartoli *et al.*, 2009). Moreover, we show that DeepCoil can be used to detect hitherto undetected coiled-coil domains in the human genome.

2 Materials and methods

2.1 Data preparation

The dataset (structures and the corresponding sequences with per-residue annotations of the coiled-coil domains) was generated by running SOCKET (cutoff 7.4 Å) (Walshaw and Woolfson, 2001) on crystallographic structures (biological assemblies) obtained from the PDB clustered to a maximum pairwise sequence identity of 50% with BLASTClust (ftp://resources.rcsb.org/sequence/clusters/bc-50.out). To increase the number of positive examples, i.e. structures containing coiled-coil domains, from each cluster, we preferentially selected structures with coiled-coil domains. To ensure the quality of the dataset, it was again filtered with CD-HIT, an accurate tool for sequence clustering, to 50% sequence identity (parameters used -c 0.5 -n 2 -T 0) and structures with a resolution <4 Å were removed. Sequences and their corresponding structures containing non-standard residues, <25 amino acid residues, or >500 amino acid residues were also removed, resulting in a final set of 21 138 entries, of which 2125 contained at least one canonical or non-canonical coiled-coil segment. For each entry, a position-specific scoring matrix (PSSM) was generated by searching the nr90 database with PSI-BLAST (Altschul *et al.*, 1997) (three iterations, 1e-3 e-value cut-off). The nr90 database was generated from the NCBI non-redundant protein sequence database (nr) using MMseqs2 (Steinegger and Söding, 2017), a tool for fast and sensitive clustering of large datasets. Since coiled coils generally contain high proportions of low-complexity regions, we did not filter out low-complexity sequences from nr90, which is generally a standard practice in the creation of filtered down sets. Entries shorter than 500 residues in the final set were randomly zero-padded from either left or right side to a constant length of 500 and one-hot encoded to generate 500 × 20 matrices (500 residues × 20 amino-acid types). Information stored in the PSSMs was zero-padded using the same procedure and encoded by transforming the values with sigmoid function, yielding matrices of size 500 × 20. The coiled-coil assignments (labels) were also zero-padded and one-hot encoded to generate 500 × 2 matrices.

2.2 Definition of training and test sets

Considering the overwhelming prevalence of the negative examples in the initial set (only ∼1, 5% residues were in coiled-coil segments), we removed half of all sequences that did not contain any coiled-coil segment. The obtained dataset was then randomly split into a training set (90% of all entries; 10 438 entries) and a test set (10% of all entries; 1193 entries; referred to as test set no. 1 henceforth), while maintaining an equal percentage of coiled-coil residues in each set (Supplementary Table S1). Furthermore, it was ensured that (i) none of the test set sequences showed more than 30% identity to any sequence of the training set and (ii) the pairwise sequence identity in the test set did not exceed 30%. A second, independent test set was derived from the test set used in a recent benchmark of coiled-coil prediction methods (Li *et al.*, 2016). Out of 1643 entries present in the test set of (Li *et al.*, 2016), we selected 518 (test set no. 2) that show no more than 30% sequence identity to any sequence of the test set no. 1 and the training set. It is important to note that these entries are not similar to the training datasets of any other coiled coil prediction method (Li *et al.*, 2016). The sequence identities were calculated using BLAST (Altschul *et al.*, 1997) (parameter -evalue 1e-2). The training set and the two test sets contained a similar proportion of parallel dimers, antiparallel dimers, trimers and tetramers. Also, the fraction of residues participating in non-canonical coiled-coil interactions is comparable to those in canonical ones (Supplementary Table S1). All the sets used in this study are available for download at https://lbs.cent.uw.edu.pl/deepcoil.

2.3 Neural network implementation and training

DeepCoil neural network was implemented in Keras (Chollet *et al.*, 2015). It consists of two, stacked convolutional layers, with 64 filters each, that scan the sequence with window sizes of 28 (first layer) and 21 (second layer). The convolutional layers are followed by a densely connected layer of 128 neurons and the output layer. 'ReLU' activation functions were used for all layers, except the output layer, where 'softmax' was used. During the training process, two dropout layers (probabilities 0.5 and 0.25, respectively) were added after each of the two convolutional layers to avoid overfitting. The training process was performed in a 5-fold cross-validation (CV) framework: the training set was divided into five equally sized parts, each containing approximately the same number of coiled-coil residues. In each CV round one part served as validation set, whereas the remaining four together as training set. The training was performed for 100 epochs with the 'Adam' (Kingma and Ba, 2014) optimizer with categorical cross-entropy as the loss function and a batch size of 64. From each CV round, a best model

(according to the F1-score) was selected and the resulting five models were used to build the final ensemble predictor. The outlined procedure was used to train two variants of the predictor, DeepCoil_SEQ, utilizing only sequence data, and DeepCoil_PSSM, utilizing sequence as well as profile data. DeepCoil_SEQ and DeepCoil_PSSM were trained using 500 × 20 (encoded sequences) and 500 × 40 matrices (encoded sequences and PSSMs), respectively. All the code necessary to replicate the analyses presented in this work is available at https://github.com/labstructbioinf/DeepCoil_Paper_2018, whereas the standalone version of DeepCoil can be downloaded from https://github.com/labstructbioinf/DeepCoil.

2.4 Identification of non-canonical coiled-coil regions

The packing in a coiled coil requires that the involved residues occupy periodically equivalent positions along the bundle interface. This cannot be achieved with undistorted helices displaying a periodicity of 3.63 residues per turn and in which the position of side-chains drift continuously. For this reason, helices in coiled coils are bent and wrapped around each other into supercoiled bundles. The handedness of the supercoiling defines whether the number of residues per turn is effectively (with respect to the bundle axis) changed to a value below 3.63 (left-handed bundles) or above 3.63 (right-handed bundles). In left-handed canonical coiled coils, the number of residues per turn is reduced to 3.5, allowing the position of the side chains to repeat after two helical turns and giving rise to a seven-residue sequence repeat (7/2 = 3.5; heptad) pattern. Non-canonical repeats are formed in an analogous way, e.g. 11/3 (hendecad), 15/4 (pentadecad) and 19/5 (nonadecad) sequence patterns are brought about by the increasingly tighter right-handed twisting of the bundle, resulting in periodicities (effective number of residues per turn) of 3.666, 3.75 and 3.8, respectively.

To identify residues that participate in non-canonical coiled-coil interactions, we analyzed 2404 coiled-coil-containing structures present in the training set and test sets nos. 1 and 2 using a modified version of the SamCC program (Dunin-Horkawicz and Lupas, 2010). The Crick parameters (Lupas and Gruber, 2005; Grigoryan and Degrado, 2011) calculated with SamCC were used to compute per-residue periodicity (P) for all residues in all coiled-coil regions using the following equation:

$$P = \frac{p}{1 - p * \Delta\omega_0/360}$$

where $p = 3.63$ is the number of residues per turn in an undistorted helix and $\Delta\omega_0$ determines the degree of twist of the coiled-coil bundle, i.e. for every residue, the angle by which the superhelix turns around the coiled-coil axis. Residues for which P deviates > 0.1 from the canonical value of 3.5 were marked as interacting in a non-canonical manner.

2.5 Human genome scanning

Sequences of human proteins were obtained from ftp://ftp.ncbi. nih.gov/genomes/Homo_sapiens/protein/protein.fa.gz and those containing non-standard amino acids or containing more than 1000 residues were removed. The remaining 94 655 sequences were used to generate PSSMs by searching the nr90 database with PSI-BLAST (three iterations, 1e-3 e-value cutoff). Sequences and corresponding PSSMs were used to predict coiled-coil regions using DeepCoil_PSSM, CCHMM_PROF and PCOILS, whereas sequences alone were used for prediction with COILS, Marcoil and Multicoil2. All predictions were performed in per-residue mode, i.e. each residue was assigned a score defining coiled-coil formation propensity. Coiled-coil regions predicted only by DeepCoil_PSSM but not by any other method were defined as sequence ranges in which (i) all per-residue scores provided by all methods except DeepCoil_PSSM were smaller than 0.5 and (ii) at least 28 consecutive residues had DeepCoil_PSSM score > 0.9. Finally, sequences containing such regions were filtered to 90% pairwise sequence identity using CD-HIT (Fu *et al.*, 2012).

3 Results and discussion

3.1 Benchmark of DeepCoil and other prediction methods

We used 10 438 structures containing 4140 uninterrupted coiled-coil regions to train DeepCoil, a neural network-based method for prediction of coiled-coil regions in protein sequences. DeepCoil was implemented in two variants: DeepCoil_SEQ that uses a single protein sequence as input, and DeepCoil_PSSM that additionally utilities evolutionary information obtained from a multiple sequence alignment. To assess the performance of the two DeepCoil variants and to compare them to the other available methods, we used two test sets: one defined in this study and the other comprising a subset of a test set used in a recent benchmark of coiled-coil prediction methods (Li *et al.*, 2016) (see Section 2). Both test sets are independent from the training set in that none of their sequences share >30% pairwise sequence identity to the sequences of the training set. The two DeepCoil variants, COILS, PCOILS, Multicoil2, Marcoil and CCHMM_PROF were used to predict coiled coils in sequences of the two test sets. To assess and quantify the accuracy of these methods in predicting the per-residue localization of coiled-coil domains (each residue of a sequence is assigned a binary value defining whether it participates in the formation of a coiled-coil region), the following metrics were used: precision, sensitivity and F1-score. Moreover, to account for partial prediction, i.e. situations where only part of a coiled coil is correctly predicted, we used the mean segment overlap (SOV) score, a measure that is based on segments rather than individual residues (Zemla *et al.*, 1999). SOV can be seen as a more detailed measure of sensitivity—its value of 0 indicates that a given coiled coil segment was missed by the predictor, whereas the value of 1 indicates a good prediction covering most of a coiled coil. In addition, we considered the percentage of coiled-coil segments having a non-zero SOV, i.e. containing at least one correctly predicted residue ('detected segments' in Tables 1 and 2). Since the performance assessment based on the aforementioned scores strongly depends on the chosen thresholds, we additionally performed receiver operating characteristic (ROC) analysis (Fig. 1) and calculated the corresponding area under ROC curves (AUC; Tables 1 and 2). The AUC scores were calculated in two variants: per-residue (see above) and per-sequence that assesses the efficiency of methods in predicting whether a given sequence contains at least a single coiled-coil region at any position (a whole sequence is assigned a binary value representing presence or absence of a coiled coil).

The benchmark performed on both test sets (Figs 1 and 2 and Tables 1 and 2) demonstrates that DeepCoil_PSSM and DeepCoil_SEQ significantly outperform the other methods. In comparison to the second best method (PCOILS_28), DeepCoil_PSSM correctly predicts at least a single residue (non-zero SOV) in nearly twice as many true coiled-coil segments (61 versus 31% and 68 versus 39% for test sets nos. 1 and 2, respectively), while providing considerably better precision (49 versus 35% and 66 versus 49%, respectively). This notion is also supported by F1 (harmonic mean

Table 1. Benchmark of DeepCoil and other coiled-coil prediction methods using a test set defined in this study (test set no.1)

Method	Profile-based[a]	Precision	Sensitivity	F1 score	Mean SOV	Detected segments	AUC[b]	AUC[c]
DeepCoil_PSSM	Yes	0.492	0.599	0.540	0.541	0.609	0.961	0.922
DeepCoil_SEQ	No	0.390	0.491	0.435	0.413	0.518	0.929	0.867
PCOILS_28[d]	Yes	0.333	0.440	0.379	0.311	0.338	0.860	0.824
PCOILS_21[d]	Yes	0.348	0.409	0.376	0.301	0.312	0.864	0.818
Marcoil	No	0.392	0.324	0.355	0.200	0.223	0.803	0.788
COILS_28[d]	No	0.386	0.317	0.348	0.199	0.218	0.803	0.806
PCOILS_14[d]	Yes	0.340	0.332	0.336	0.269	0.307	0.850	0.795
Multicoil2	No	0.550	0.242	0.336	0.124	0.127	0.732	0.691
COILS_21[d]	No	0.332	0.322	0.327	0.228	0.252	0.836	0.794
CCHMM_PROF	Yes	0.255	0.310	0.280	0.234	0.237	—[e]	0.803
COILS_14[d]	No	0.252	0.306	0.276	0.261	0.309	0.832	0.760

Note: The methods are ordered according to the decreasing F1 score. The corresponding ROC plots are shown in Figures 1A and 2A.
[a]Indication whether a method is profile- or sequence-based.
[b]AUC scores for per-residue classification.
[c]AUC scores for per-sequence classification.
[d]Suffix refers to the size of the scanning window.
[e]CCHMM_PROF does not return per-residue probabilities.

Table 2. Benchmark of DeepCoil and other coiled-coil prediction methods using test set no. 2 (subset of the dataset defined in the study of Li *et al.*)

Method	Profile-based[a]	Precision	Sensitivity	F1 score	Mean SOV	Detected segments	AUC[b]	AUC[c]
DeepCoil_PSSM	Yes	0.662	0.628	0.645	0.609	0.678	0.946	0.954
DeepCoil_SEQ	No	0.607	0.465	0.527	0.424	0.538	0.903	0.897
PCOILS_28[d]	Yes	0.495	0.449	0.471	0.359	0.388	0.837	0.833 (0.800*)
PCOILS_21[(a)]	Yes	0.516	0.387	0.442	0.331	0.360	0.835	0.822 (0.800*)
COILS_28[d]	No	0.559	0.278	0.372	0.219	0.239	0.788	0.823 (0.783*)
PCOILS_14[d]	Yes	0.526	0.283	0.368	0.271	0.323	0.816	0.810 (0.800*)
Marcoil	No	0.540	0.263	0.354	0.209	0.227	0.797	0.830 (0.808*)
COILS_21[d]	No	0.512	0.266	0.350	0.232	0.259	0.816	0.813 (0.783*)
COILS_14[d]	No	0.446	0.260	0.328	0.255	0.311	0.804	0.768 (0.783*)
CCHMM_PROF	Yes	0.445	0.257	0.326	0.192	0.192	—[e]	0.848 (0.811*)
Multicoil2	No	0.610	0.170	0.265	0.098	0.101	0.695	0.706 (0.699*)

Note: The methods are ordered according to the decreasing F1 score. The corresponding ROC plots are shown in Figures 1B and 2B.
*Per-sequence AUC scores obtained from a work of Li *et al.*; for the description of the remaining footnotes see Table 1.

of precision and sensitivity) and AUC scores. The AUC scores were calculated for two prediction tasks, i.e. per-residue and per-sequence detection of coiled-coil domains. In both cases, DeepCoil performs substantially better than any other method. The per-sequence AUC scores obtained in this study are in good agreement ($\rho = 0.93$, P-value < 0.01) with those obtained by Li *et al.* (2016) (Table 2), supporting the validity of our benchmarking procedure. It is also important to note that DeepCoil_SEQ, a variant of DeepCoil that does not require the computationally intensive task of profile calculation, is better than the currently available profile-based methods. Finally, the CV analysis (see Section 2) and benchmarks on the test sets resulted in very similar statistics, indicating that model is not overfitted, e.g. due to the imbalance between the training set (max. 50% pairwise sequence identity) and the test sets (max. 30% pairwise sequence identity).

To assess the performance of DeepCoil in the prediction of various types of coiled-coil domains, the test sets were divided into subsets, each of which contained parallel dimers, antiparallel dimers, trimers and tetramers. In addition, a test set variant containing only coiled-coil residues interacting in non-canonical fashion was generated. All these sets were used to benchmark DeepCoil and the other methods (Supplementary Figs S1 and S2 and Supplementary Tables S2 and S3). The obtained results indicate that despite a bias in the training set (Supplementary Table S1), which for example contains only a tiny number of non-canonical coiled coils, DeepCoil performs equally well in the prediction of both common and rare coiled coil types.

3.2 Identification of coiled-coil regions in the human genome

In the test set no. 1, we identified 57 coiled-coil segments with a length of 14 or more residues, all of which are correctly assigned by DeepCoil, but not by any other method. This observation prompted us to check whether DeepCoil can be used to detect previously undetected coiled-coil domains in the human genome. To this end, we scanned ~100.000 human proteins using methods listed in Tables 1 and 2, and identified those that contain coiled-coil segments detected only by DeepCoil (see Section 2 for details). We found 35 coiled-coil regions that were not predicted by the other methods

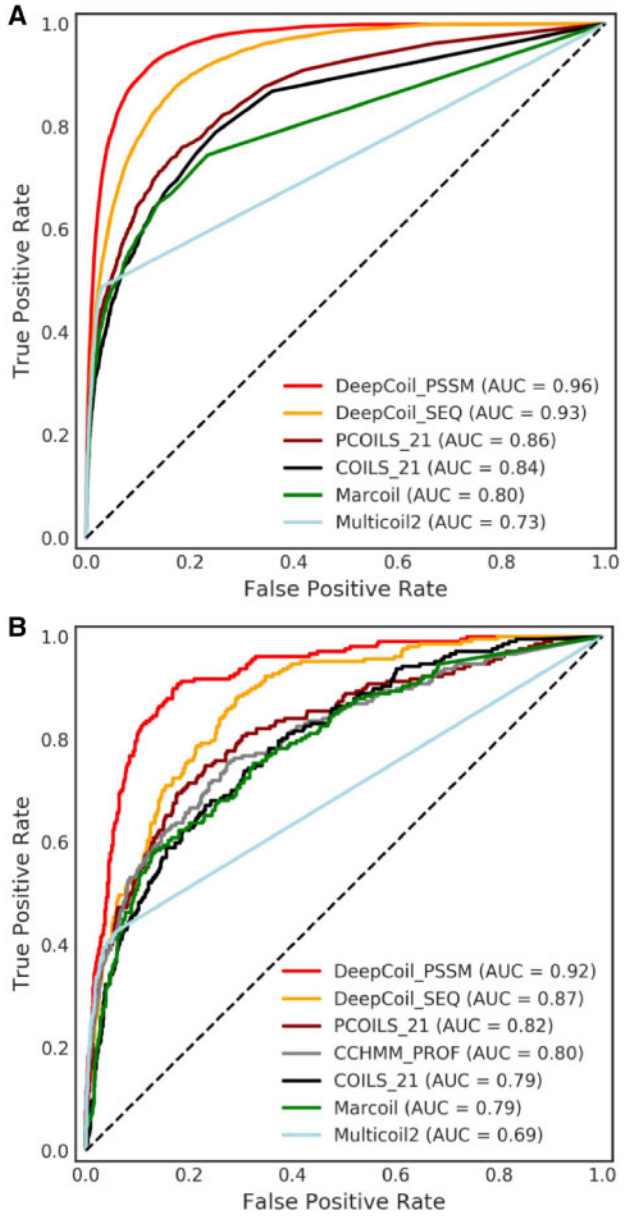

Fig. 1. ROC analysis performed based on test set no. 1 (defined in this study). Panels **(A)** and **(B)** denote the per-residue and per-sequence predictions, respectively. For clarity, some variants of PCOILS and COILS were omitted. For all AUC scores see Tables 1 and 2

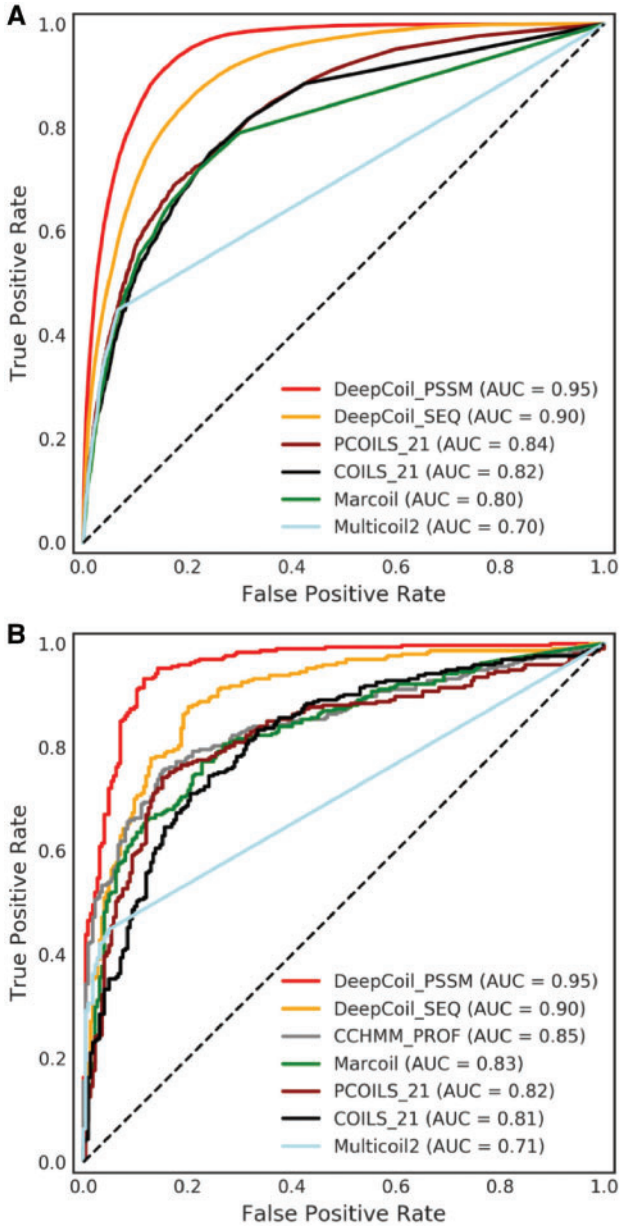

Fig. 2. ROC analysis performed based on test set no. 2 (defined based on the work of Li *et al.*). Panels **(A)** and **(B)** denote the per-residue and per-sequence predictions, respectively. For clarity, some variants of PCOILS and COILS were omitted. For all AUC scores see Tables 1 and 2

even at a permissive cutoff of 0.5 but were predicted by DeepCoil at a very strict cut-off of 0.9 (Supplementary Material S1). For ten of these predicted coiled coils, there are available experimental structures. In nine of them, we indeed found a coiled-coil structure and only one predicted coiled-coil corresponds to a regular helical region. Importantly, only four of these ten structures are present in the training set. For the other 16 predicted coiled-coil regions, we identified homologs with known experimental structures using HHpred (Remmert *et al.*, 2011). Manual inspection of these structures revealed that in 14 cases the segments predicted by DeepCoil form coiled coils in the homologous structures and only in 2 cases they don't do so. Finally, for the remaining nine predictions, we could not identify homologs of known structure, but considering the above results, it is highly probable that most of them correspond to hitherto uncharacterized coiled coils.

4 Conclusions

Our results show that DeepCoil outperforms the current best methods for the prediction of canonical and non-canonical coiled coils, making it an attractive choice for genome-wide scans of previously uncharacterized coiled-coil domains. DeepCoil is available as a standalone method at https://github.com/labstructbioinf/DeepCoil and as an easy-to-use web server at https://toolkit.tuebingen.mpg.de/#/tools/deepcoil (Zimmermann *et al.*, 2018).

Funding

This work was supported by the Polish National Science Centre [grant number 2015/18/E/NZ1/00689 to S.D.H.]. V.A. was supported by institutional funds from the Max Planck Society. Computations were carried out with the

support of the Interdisciplinary Centre for Mathematical and Computational Modeling (ICM) University of Warsaw [grant number GA71-24 to S.D.H.].

Conflict of Interest: none declared.

References

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Alvarez,B.H. *et al*. (2010) A transition from strong right-handed to canonical left-handed supercoiling in a conserved coiled-coil segment of trimeric autotransporter adhesins. *J. Struct. Biol*., **170**, 236–245.

Armstrong,C.T. *et al*. (2011) SCORER 2.0: an algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences. *Bioinformatics*, **27**, 1908–1914.

Bartoli,L. *et al*. (2009) CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information. *Bioinformatics*, **25**, 2757–2763.

Chollet,F. *et al*. (2015) Keras. https://github.io.

Delorenzi,M. and Speed,T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, **18**, 617–625.

Dunin-Horkawicz,S. and Lupas,A.N. (2010) Measuring the conformational space of square four-helical bundles with the program samCC. *J. Struct. Biol*., **170**, 226–235.

Fu,L. *et al*. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Grigoryan,G. and Degrado,W.F. (2011) Probing designability via a generalized model of helical bundle geometry. *J. Mol. Biol*., **405**, 1079–1100.

Gruber,M. *et al*. (2005) REPPER–repeats and their periodicities in fibrous proteins. *Nucleic Acids Res*., **33**, W239–W243.

Gruber,M. *et al*. (2006) Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol*., **155**, 140–145.

Kingma,D.P. and Ba,J.L. (2014) Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li,C. *et al*. (2015) Computational characterization of parallel dimeric and trimeric coiled-coils using effective amino acid indices. *Mol. Biosyst*., **11**, 354–360.

Li,C. *et al*. (2016) Critical evaluation of in silico methods for prediction of coiled-coil domains in proteins. *Brief. Bioinform*., **17**, 270–282.

Lupas,A. *et al*. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.

Lupas,A.N. *et al*. (2017) The structure and topology of alpha-helical coiled coils. *Subcell. Biochem*., **82**, 95–129.

Lupas,A.N. and Bassler,J. (2017) Coiled coils - a model system for the 21st century. *Trends Biochem. Sci*., **42**, 130–140.

Lupas,A.N. and Gruber,M. (2005) The structure of alpha-helical coiled coils. *Adv. Protein Chem*., **70**, 37–78.

McFarlane,A.A. *et al*. (2009) The use of coiled-coil proteins in drug delivery systems. *Eur. J. Pharmacol*., **625**, 101–107.

Remmert,M. *et al*. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol*., **35**, 1026–1028.

Szczepaniak,K. *et al*. (2018) Variability of the core geometry in parallel coiled-coil bundles. *J. Struct. Biol*., **204**, 117–124.

Trigg,J. *et al*. (2011) Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. *PLoS One*, **6**, e23519.

Vincent,T.L. *et al*. (2013) LOGICOIL–multi-state prediction of coiled-coil oligomeric state. *Bioinformatics*, **29**, 69–76.

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.

Woolfson,D.N. (2017) Coiled-Coil design: updated and upgraded. *Subcell. Biochem*., **82**, 35–61.

Zemla,A. *et al*. (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.

Zimmermann,L. *et al*. (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core. *J. Mol. Biol*., **430**, 2237–2243.