PrOCoil - Advances in predicting two- and three-stranded coiled coils

Carsten C. Mahrenholz*², Ingrid G. Abfalter*¹, Ulrich Bodenhofer¹, Rudolf Volkmer² and Sepp Hochreiter¹

¹Institute of Bioinformatics, Johannes Kepler University, Linz, Austria
²Institute of Medical Immunology, Charité Medical School, Berlin, Germany

Abstract The aim of this project was to investigate which specific properties influence the oligomerization of coiled coil (CC) proteins. Based on comprehensive peptide libraries of GCN4 mutants, our results indicated that the formation of oligomeric structures strongly depends on various specific positions of a heptad. To investigate these dependencies, we collected all known dimeric and trimeric CCs in the PDB and applied Support Vector Machines (SVMs) to classify them. We used pairwise feature extraction to identify characteristic amino acid patterns that determine the rules for oligomerization. These rules form the basis of our classification tool *PrOCoil*, which also visualizes the contribution of each individual amino acid to the overall oligomeric tendency of a given coiled coil sequence. *PrOCoil* is also available as a web tool (<http://www.bioinf.jku.at/software/procoil/>).

Introduction CCs consist of 2 to 5 alpha-helices that are wrapped around each other with a left-handed super-helical twist. Their ubiquity and the stable interactions of their helices make coiled coils ideal building blocks for **designing novel proteins**. Furthermore, CC interactions have recently attracted attention as **promising drug targets**.

Peptide libraries As multiple substitutions can cause a switch from dimeric to trimeric or higher structures, we used a peptide-array approach to investigate the effect of double-substitutions at positions *a/d*, *a/e*, and *d/g* of the GCN4 leucine-zipper (**C**). Several GCN4 mutants were selected and tested for homo- or heterospecific interaction, resulting in a network of CC interactions (**D**). The results were verified by means of analytical ultra centrifugation. [2]

Schematic representation and structure of the parallel dimeric coiled-coil motif [1] (A) Helical wheel diagram looking down the helix axis from the N- to the C-terminus. Heptad positions are labelled a-g and a'-g' respectively. Positions a, d, e and g are colour-coded. (B) In the side view, the helical backbones are represented by cylinders, the side chains by knobs, and the path of the polypeptide chain is indicated by a line wrapped around the cylinders. For simplicity, the supercoiling of the helices is not shown. While residues at positions a (purple) and d (blue) make up the hydrophobic interface, residues at positions e (orange) and g (red) pack against the hydrophobic core. They can participate in interhelical electrostatic interactions between residue *i* (g position) of one helix and residue *i*+5 of the other helix (e' position, belonging to the next heptad), as indicated by the hatched bars.

Based on these findings, bioinformatics tools were used to understand the described results.

Data preparation (E) By scanning the whole PDB with the tool SOCKET [3], we retrieved all coiled coil sequences with known 3D structure (**F**). After removing identical (sub)sequences, we created a database of 385 dimeric and 92 trimeric coiled coil sequences with heptad registers (abcdefg)_n assigned by SOCKET. To augment this set with newly sequenced genome data, we employed a sophisticated BLAST approach with stringent filtering, which resulted in a combined **dataset of 2043 dimers and 791 trimers**. In contrast to hitherto published approaches to coiled coil analysis, we clustered our data to minimize statistical bias. Therefore, conclusions based on the dataset are representative of coiled coil sequences in general and are not biased by the variability of a few sequences that are in the limelight of scientific interest and thus more prevalent in the PDB. We used both the PDB and the augmented dataset to create clustered datasets with a 60% sequence identity threshold.

Coiled Coil Sequences in the PDB

	parallel	anti-parallel	
DIMER			
homo	225	213	438
hetero	976 (488)	7614 (3807)	8590
	1201	8052	
TRIMER			
homo	172	5	177
hetero	246 (82)	1098 (366)	1344
	418	1103

PrOCoil output example: Sequence profiling plot visualizing the weight and oligomeric tendency of each amino acid position for the (G) homodimeric cJun-cJun complex and the (H) trimeric hemagglutinin of influenza A. **Overview of (I)** model selection results obtained by nested cross-validation and (K) by cross-validation on the entire data set. (L) Schematic view of pairwise feature-based heptade-spanning networks that govern oligomerization. Figures were created using PyMOL (<http://pymol.sourceforge.net/>).

Profiling using PrOCoil (G, H) Using the coiled coil kernel, an SVM generated rules by optimizing the pattern weights such that the combined rules achieved maximum discrimination between dimers and trimers. We then used this information to implement a prediction and sequence profiling tool, *PrOCoil*, that characterizes the overall oligomeric tendency of a coiled coil sequence by displaying each amino acid's contribution to the rules in which it participates in a specific sequence. The figures above depict the sequence profiling plots of a typical dimer (c-Jun, PDB entry 1jun) and a typical trimer (hemagglutinin, PDB entry 1htm) generated using *PrOCoil*. Oligomeric tendencies are indicated by the areas above and below the base line, which equate respectively to the positive/trimeric contributions and the negative/dimeric contributions of the involved patterns.

Classification Results (I, K) To identify the SVM classifier (i.e., model) with the optimum SVM and coiled coil kernel parameters, we had to assess which model performs best on future (previously unseen) data. For this purpose, we applied nested cross-validation: The best model, classified test (i.e., unknown) sequences with **88.6% accuracy**, even though they had only a maximum identity of 60% to any (known) coiled coil with which the SVM was trained.

Conclusion Our results indicate that the two-amino acid dependencies found in our biochemical experiments (C) are parts of the equation that describes CC-oligomerization. We demonstrated that complex networks (L) of dependencies govern oligomerization and provided a tool that classifies CC sequences with outstanding accuracy.

[1] Mason *et al.* Biochem. Soc. Trans. 36, 1442-1447 (2008).
[2] Portwich *et al.* Angewandte Chemie Int. Ed. Engl. 46, 1654-1657 (2007).
[3] Walshaw *et al.* J. Mol. Biol. 307, 1427-50 (2001).
[4] <http://scienceblogs.com/mimiprings/129611>