SCALABLE COMPUTING: PRACTICE AND EXPERIENCE
Volume 8, Number 2, pp. 167–171. http://www.scpe.org
ISSN 1895-1767

GENOME-WIDE IDENTIFICATION AND COMPARATIVE ANALYSIS OF COILED-COIL PROTEINS

ANNKATRIN ROSE*, ERIC A. STAHLBERG†, AND IRIS MEIER‡

Abstract. The α-helical coiled-coil is a protein structure motif well suited for computational prediction. To study the occurrence of coiled-coil proteins throughout different kingdoms, we have computationally identified and clustered long coiled-coil proteins from 23 fully-sequenced genomes. Our results indicate that long coiled-coil proteins occur with higher frequency in eukaryotes than prokaryotes, with kingdom-specific families observed in plants and animals. We have established searchable protein databases containing prediction data for *Arabidopsis*, rice and *Chlamydomonas* coiled-coil proteins to facilitate further studies.

Key words. coiled-coil, Multicoil, protein structure prediction, clustering, database

1. Introduction. Coiled-coil proteins play an important role in the spatial and temporal organization of cellular processes, such as signal transduction, cell division, structural integrity and motility. Long coiled-coil domains serve as "cellular velcro" and form dynamic fibers and scaffolds, allowing proteins to act as molecular "zippers", adapters, spacers, and motors in macro-molecular structures [1]. These biophysical properties qualify coiled-coil proteins as candidates for nanotechnology applications and biosensors [2], [3], [4], [5]. Mutations in coiled-coil proteins have been implicated in a growing number of human diseases from muscular dystrophies and neurodegenerative diseases to premature aging syndromes and cancer, illustrating their importance in a biological context [6], [7], [8]. In contrast to animals and yeast, only a handful of long coiled-coil proteins have been studied in plants and prokaryotes.

The coiled-coil motif consists of two or more α-helices winding around each other in a supercoil [9] often serving as a protein oligomerization domain. It is characterized by a heptad repeat in the primary sequence, which facilitates computational prediction of coiled-coil domains [10]. The most commonly used prediction programs include COILS, the PairCoil/MultiCoil algorithms, the hidden Markov model-based Marcoil, and PCOILS—an improved version of COILS using profiles [11], [12], [13], [14], [15], [16]. Using computational predictions, it has been estimated that approximately 10% of all proteins in an organism contain coiled-coil sequences [17]. Taking advantage of the availability of fully-sequenced genomes, it is now possible to conduct comprehensive computational analyses and comparisons of the coiled-coil protein composition of different organisms [18], [19].

2. Identification and Selection of Long Coiled-Coil Proteins. Using the coiled-coil prediction program MultiCoil [13], we have identified all long coiled-coil proteins from 23 fully-sequenced genomes. The MultiCoil program was downloaded from http://theory.lcs.mit.edu/multicoil and installed on the Ohio Supercomputer Center 512 processor Pentium 4 cluster. Whole genome sequence files were downloaded from the European Bioinformatics Institute proteome analysis database (http://www.ebi.ac.uk/proteome/) and processed as shown in Figure 2.1. Coiled-coil prediction raw output was generated by running the sequences through the locally installed MultiCoil program using a cutoff score of 0.5 and window size of 28. A Java-based program suite, ExtractProp, was developed to post-process the raw output by ignoring small gaps (less than 25 residues) and setting a minimum domain length of 20 residues to allow for the formation of a stable helix in the secondary structure of the protein (see Figure 2.2 for an example).

To identify coiled-coil proteins putatively involved in structural functions in the cell, the ExtractProp suite further selected for proteins containing at least one domain of at least 70 amino acids, two domains of at least 50 amino acids, or three or more domains of at least 30 amino acids in length ("long coiled-coils" in Figure 3.1). The ExtractProp suite is available for download at http://www.osc.edu/research/bioinformatics/software.shtml.

3. Results of Coiled-Coil Prediction. In contrast to older studies which predicted 10% coiled-coil sequences [17], we took a more restrictive approach to predicting coiled-coils by introducing a minimum domain length cutoff to eliminate short sequences unlikely to form stable structures. We find on average 6.4% of eukaryotic proteins and 3.5% of prokaryotic proteins are predicted to contain coiled-coil structures (also see Figure 3.1, top panel). Long coiled-coil domains were found underrepresented in most bacterial genomes; however, both archaea and eukaryotes contain longer coiled-coil domains than eubacteria. This result was especially pronounced for longer coiled-coils more than 250 amino acids in length (see Figure 3.1, bottom panel).

*Department of Biology, Appalachian State University, 572 Rivers Street, Boone, NC 28608, USA
†Ohio Supercomputer Center, 1224 Kinnear Road, Columbus, OH 43212, USA
‡Department of Plant Biology and Plant Biotechnology Center, Ohio State University, 1060 Carmack Road, Columbus, OH 43210, USA. This work was supported by the National Science Foundation 2010 Project (grant no. NSF 0209339 to I.M.)

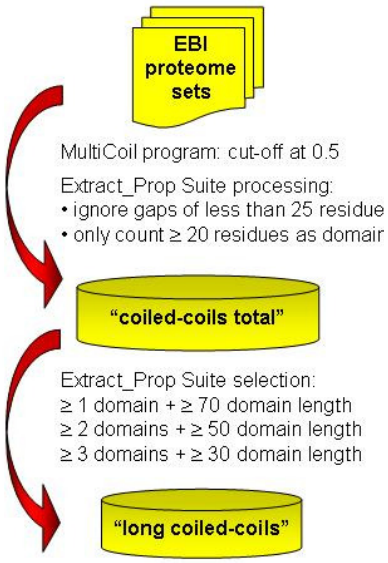

FIG. 2.1. *Sequence processing through MultiCoil and ExtractProp.*

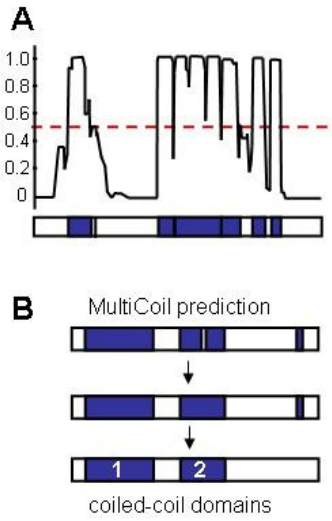

FIG. 2.2. *Coiled-coil structure prediction by MultiCoil and ExtractProp processing. (A) MultiCoil prediction (scores per residue) of FPP1 protein sequence with score cutoff of 0.5 (red line) and predicted coiled-coil domains shown in blue. (B) ExtractProp processing of MultiCoil raw output to eliminate short gaps and stretches of predicted coiled-coil too short to form a stable helix.*

4. Clustering of Coiled-Coil Protein Sequences. Due to the characteristic sequence repeat pattern arising from the structural constraints of the coiled-coil motif, sequences predicted to form coiled-coils often interfere with the statistical determination of significant sequence similarities. To circumvent this problem, we developed a sequence comparison and clustering strategy based on masking the identified coiled-coil domains to eliminate similarities based on structural constraints of the coiled-coil (see Figure 4.1). Using this method, we compared and grouped all identified long coiled-coil proteins based on sequence similarities outside their coiled-coil regions.

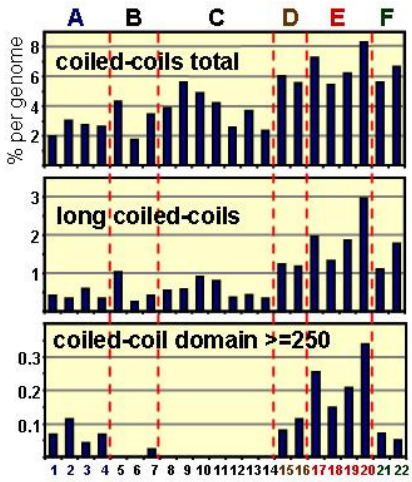

FIG. 3.1. *Coiled-coil proteins predicted per genome (as percent of all protein sequences). (A) archaea: 1, Thermoplasma acidophilum, 2, Methanococcus jannaschii, 3, Archeoglobus fulgidus, 4, Sulfolobus solfataricus; (B) gram-positive bacteria, 5, Mycoplasma genitalium, 6, Mycobacterium tuberculosis, 7, Bacillus subtilis; (C) gram-negative bacteria, 8, Clamydia pneumoniae, 9, Heliobacter pylori, 10, Borrelia burgdorferi, 11, Synechocystis sp. PCC6803, 12, Escherichia coli, 13, Chromobacterium violaceum, 14, Agrobacterium tumefaciens; (D) yeasts, 15, Schizosaccharomyces pombe, 16, Saccharomyces cerevisiae; (E) metazoa, 17, Drosophila melanogaster, 18, Caenorhabditis elegans, 19, Mus musculus, 20, Homo sapiens; F, plants, 21, Arabidopsis thaliana, 22, Oryza sativa.*

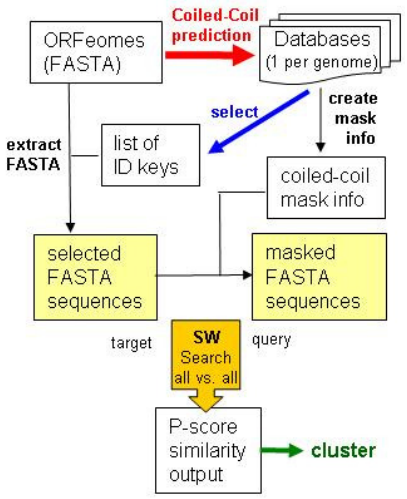

FIG. 4.1. *Flowchart for clustering analysis.*

First, coiled-coil sequences were masked by replacing amino acids predicted to be inside a coiled-coil region with the generic letter X. The masked sequence set was then compared in an all-against-all approach using the Smith-Waterman (SW Search) sequence comparison algorithm [20]. Smith-Waterman analysis was accomplished with the TimeLogic DeCypher system using the blossum62 scoring matrix. Extraction of Smith-Waterman scores and distances was done using the OSC Apple G5 cluster and elements from the ExtractProp software suite. Clustering of sequence results was done by grouping sequences using a modified version of Kruskal's minimum cost spanning tree algorithm [21] as described in [19]. The threshold criteria for determining cluster inclusion were at 1.0e-15 for Smith-Waterman P-score similarity between sequences with coiled-coil regions masked.

5. Clustering Results. During clustering of the predicted long coiled-coil proteins from all analyzed genomes, several kingdom-specific coiled-coil protein families emerged. The structural maintenance of chromosomes (SMC) proteins and their relatives stood out as the only long coiled-coil protein family conserved throughout all kingdoms. Motor

proteins, such as myosin and kinesins, as well as membrane tethering and vesicle transport proteins are the dominant eukaryotic long coiled-coil proteins. A number of plant proteins with unknown function could be grouped with already characterized animal and yeast proteins in these families. A group of coiled-coil proteins present in animals but apparently absent in plants and yeast are nuclear matrix and intermediate filament proteins, such as the nuclear lamins, as well as membrane-cytoskeleton cross-linkers and scaffolding proteins. Metazoan mitotic motility proteins and microtubule organization center components also lack homologs in plants, consistent with known differences in mitotic microtubule nucleation between these kingdoms (see Table 5.1 and [19]).

While masking the coiled-coil domains before sequence comparison significantly increased the specificity of the clustering analysis, the method has limitations regarding protein sequences with high coiled-coil content. These were not included due to insufficient sequence left after masking the coiled-coil residues to provide significant P-scores during Smith-Waterman comparison.

6. Database Development. The selected long coiled-coil proteins from the *Arabidopsis*, rice, and *Chlamydomonas* genome were used to build databases utilizing MySQL version 4.1 connected through JDBC to the searchable website (`http://www.coiled-coil.org/` [18]). The *Arabidopsis* coiled-coil protein database ARABI-COIL integrates information on number, size, and position of predicted coiled-coil domains with subcellular localization signals, transmembrane domains, and available functional annotations (`http://www.coiled-coil.org/arabidopsis/`). The development of a corresponding rice coiled-coil protein database is in progress (`http://www.coiled-coil/org/rice/`), which will allow for comparative analysis of long coiled-coil proteins encoded by different plant genomes. We are in the process of adding coiled-coil prediction data for the green algae *Chlamydomonas reinhardtii* in collaboration with the *Chlamydomonas* genome project (`http://genome.jgi-psf.org/Chlre3/Chlre3.home.html`). The results from the clustering analysis will be integrated to improve the annotation of so far uncharacterized plant coiled-coil proteins in these databases.

7. Conclusions and Outlook. Long coiled-coil proteins are predominantly involved in subcellular infrastructure maintenance and trafficking control. Many of these proteins seem to be missing in plants. Due to the difficulties identifying plant homologs of many metazoan coiled-coil proteins, functional studies will have to reveal whether so far uncharacterized plant proteins fulfill functions similar to metazoan counterparts. The generated coiled-coil protein databases can now serve as a data-mining tool to sort and browse plant long coiled-coil proteins, therefore facilitating the identification and selection of candidate proteins of interest. Using the ARABI-COIL database, we identified putative *Arabidopsis* membrane-bound, nuclear, and organellar long coiled-coil proteins for ongoing experimental studies.

TABLE 7.1
Functional groups of coiled-coil proteins identified through clustering and their representation in different kingdoms.

Protein Function	**Species**
Chromatin organization and maintenance, DNA repair	all kingdoms
Transcription and translation	all kingdoms
Protein trafficking and quality control	prokaryotes and organelles
Membrane channels and regulation of influx/export	prokaryotes
Sensors and signal transduction	eukaryotes
Membrane organization, stabilization, and dynamics	eukaryotes
Cell adherence	eukaryotes and parasitic prokaryotes
Mechanical fiber and meshwork formation	eukaryotes
Motility	eukaryotes
Cytoskeleton organization, stabilization, and dynamics	eukaryotes, predominantly metazoa
Mitotic spindle assembly and checkpoint control	metazoa and yeast

REFERENCES

[1] A. ROSE AND I. MEIER, *Scaffolds, levers, rods and springs: diverse cellular functions of long coiled-coil proteins*, Cellular and Molecular Life Sciences 61:1996-2009, 2004.

[2] H. CHAO, D. L. BAUTISTA, J. LITOWSKI, R. T. IRVIN AND R. S. HODGES, *Use of a heterodimeric coiled-coil system for biosensor application and affinity purification*, Journal of Chromatography. B, Biomedical Sciences and Applications, 715:307-329, 1998.

[3] A. J. DOERR AND G. L. MCLENDON, *Design, folding, and activities of metal-assembled coiled coil proteins*, Inorganic Chemistry, 43:7916-7925, 2004.

[4] R. R. NAIK, S. M. KIRKPATRICK AND M. O. STONE, *The thermostability of an α-helical coiled-coil protein and its potential use in sensor applications*, Biosensors and Bioelectronics, 16:1051-1057, 2001.
[5] M. M. STEVENS, S. ALLEN, J. K. SAKATA, M. C. DAVIES, C. J. ROBERTS, S. J. B. TENDLER, D. A. TIRRELL AND P. M. WILLIAMS, *pH-dependent behavior of surface-immobilized artificial leucine zipper proteins*, Langmuir: The ACS Journal of Surfaces and Colloids, 20:7747-7752, 2004.
[6] L. MOUNKES, S. KOZLOV, B. BURKE AND C. L. STEWART, *The laminopathies: nuclear structure meets disease*, Current Opinion in Genetics and Development, 13:223-230, 2003.
[7] L. C. MOUNKES AND C. L. STEWART, *Aging and nuclear organization: lamins and progeria*, Current Opinion in Cell Biology, 16:322-327, 2004.
[8] T. M. MAGIN, J. REICHELT AND M. HATZFELD, *Emerging functions: diseases and animal models reshape our view of the cytoskeleton*, Experimental Cell Research, 301:91-102, 2004.
[9] P. BURKHARD, J. STETEFELD AND S. V. STRELKOV, *Coiled coils: a highly versatile protein folding motif*, Trends in Cell Biology, 11:82-88, 2001.
[10] D. A. PARRY, *Coiled-coils in α-helix-containing proteins: analysis of residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins*, Bioscience Reports, 2:54-63, 1982.
[11] A. LUPAS, M. VAN DYKE AND J. STOCK, *Predicting coiled coils from protein sequences*, Science, 252:1162-1164, 1991.
[12] B. BERGER, D. B. WILSON, E. WOLF, T. TONCHEV, M. MILLER AND P. S. KIM, *Predicting coiled coils by use of pairwise residue correlations*, Proceedings of the National Academy of Sciences U.S.A., 92:8259-8263, 1995.
[13] E. WOLF, P. S. KIM AND B. BERGER, *MultiCoil: a program for predicting two- and three-stranded coiled coils*, Protein Science, 6:1179-1189, 1997.
[14] M. DELORENZI AND T. SPEED, *An HMM model for coiled-coil domains and a comparison with PSSM-based predictions*, Bioinformatics, 18:617-625, 2002.
[15] M. GRUBER, J. SÖDING AND A. N. LUPAS, *REPPER-repeats and their periodicities in fibrous proteins*, Nucleic Acids Research, 33:239-243, 2005.
[16] M. GRUBER, J. SÖDING AND A. N. LUPAS, *Comparative analysis of coiled-coil prediction methods*, Journal of Structural Biology, 155:140-145, 2006.
[17] J. LIU AND B. ROST, *Comparing function and structure between entire genomes*, Protein Science, 10:1970-1979, 2001.
[18] A. ROSE, S. MANIKANTAN, S. J. SCHRAEGLE, M. A. MALOY, E. A. STAHLBERG AND I. MEIER, *Genome-wide identification of Arabidopsis coiled-coil proteins and establishment of the ARABI-COIL database*, Plant Physiology, 134:927-939, 2004.
[19] A. ROSE, S. J. SCHRAEGLE, E. A. STAHLBERG AND I. MEIER, *Coiled-coil composition of 22 proteomes—differences and common themes in subcellular infrastructure and traffic control*, BMC Evolutionary Biology, 5:66, 2005.
[20] T. F. SMITH AND M. S. WATERMAN, *Identification of common molecular subsequences*, Journal of Molecular Biology, 147:195-197, 1981.
[21] J. B. KRUSKAL, *On the shortest spanning subtree of a graph and the traveling salesman problem*, Proceedings of the American Mathematical Society, 7:48-50, 1956.

Edited by: Dazhang Gu
Received: Jan 16, 2007
Accepted: April 15, 2007