

## Genome-wide identification and comparative analysis of coiled-coil proteins

Annkatrin Rose<sup>1</sup>, Eric A. Stahlberg<sup>2</sup>, and Iris Meier<sup>1</sup>

<sup>1</sup> Department of Plant Biology and Plant Biotechnology Center, Ohio State University, 1060 Carmack Road, Columbus, OH 43210

<sup>2</sup> Ohio Supercomputer Center, 1224 Kinnear Road, Columbus, OH 43212, USA

Long 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 form dynamic fibers and scaffolds, allowing proteins to act as molecular "zippers", adapters, spacers, and motors in macro-molecular structures. The coiled-coil motif consists of two or more alpha-helices winding around each other in a supercoil and is characterized by a heptad repeat in the primary sequence, which facilitates computational prediction of coiled-coil domains.

Using the coiled-coil prediction program MultiCoil, we have identified all long coiled-coil proteins from 23 fully-sequenced genomes. Due to the characteristic sequence repeat pattern of the coiled-coil motif, regions predicted to form coiled-coils often interfere with the statistical determination of significant sequence homologies. 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. Using this method, we compared and grouped all identified long coiled-coil proteins based on sequence similarities outside their coiled-coil regions.

Long coiled-coil domains were found underrepresented in most bacterial genomes, however both archaea and eukaryotes contain longer coiled-coil domains than bacteria. Several clusters of kingdom-specific coiled-coil protein families emerged, while at the same time a number of plant proteins with unknown function could be grouped with already characterized animal and yeast proteins. The structural maintenance of chromosomes proteins and their relatives stood out as the only long coiled-coil protein family conserved throughout all kingdoms. Motor proteins as well as membrane tethering and vesicle transport proteins are the dominant eukaryotic long coiled-coil proteins.

We have built a searchable *Arabidopsis thaliana* coiled-coil protein database, ARABI-COIL [<http://www.coiled-coil.org/arabidopsis/>], integrating information on number, size, and position of predicted coiled-coil domains with subcellular localization signals, transmembrane domains, and available functional annotations. ARABI-COIL serves as a data-mining tool to sort and browse *Arabidopsis* long coiled-coil proteins, therefore facilitating the identification and selection of candidate proteins of interest for specific research areas. Using the database, we identified putative *Arabidopsis* membrane-bound, nuclear, and organellar long coiled-coil proteins. The development of a corresponding rice coiled-coil protein database is in progress, which will allow for comparative analysis of long coiled-coil proteins encoded by different plant genomes. The results from the clustering analysis will be integrated to improve the annotation of so far uncharacterized plant coiled-coil proteins in these databases.