Unveiling Substructure Similarity: Mining of Geometrically Conserved Structural Motifs and Long-Range Contacts in Protein Structure Data

Florian Kaiser[1,2] and Dirk Labudde[1,*]

[1]University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany, and

[2]Biotechnology Center (BIOTEC), Technische Universität Dresden, Tatzberg 47-49, 01307 Dresden, Germany.

[*]Corresponding author: dirk.labudde@hs-mittweida.de

Small conserved structural units in proteins play key roles for protein function and were extensively researched. Responsible for nucleotide interaction [1], ion fixation [2], or catalytic activity [3], such structural motifs are indispensable to bridge the gap between protein structure and function. Consequently, computational methods to screen for known structural motifs were successfully applied to predict protein function [4] or to identify remote homologous proteins [5].

Due to often noncontiguous residues in protein sequence, structural motif identification by multiple sequence alignment (MSA) techniques can be a delicate task [6, 7]. However, the *a priori* knowledge of a reference structural motif is vital to screen for this template in target structure data to infer common function, fold, or ancestry. Unfortunately, the discovery of geometrically conserved structural motifs is still a major hurdle. To address this problem we present a unsupervised method to identify structural motifs in arbitrary-sized sets of protein structures. Based on the extension of earlier work [8] we included geometrical evaluation to carve out highly similar properties of potential motifs, such as congruent side chain orientations. Additionally, our method is eligible to spot conserved long-range structure contacts, often difficult to identify by MSA.

By using methods originated from data mining, we were able to automatically pinpoint catalytic and ion binding sites in defined protein families as a proof-of-concept. We call these patterns family-specific structural motifs. Additionally, the application to over 13,000 nonredundant protein structures indicates the prevalence of ubiquitous structural motifs that can be observed throughout the protein universe. The existence of such molecular building blocks – independent of any specific fold or function – is supported by previous findings [9].

The developed concepts are not limited to protein structures and will be applied on other macromolecular structure data (e.g. DNA/RNA structures, protein-nucleotide or protein-ligand complexes) in near future. We envision that our approach is suitable to derive fingerprint libraries of naturally observed structural motifs, which could be used to assess protein family association or to predict protein function.

References

[1] R. B. Darnell. Developing global insight into RNA regulation. *Cold Spring Harb. Symp. Quant. Biol.*, 71:321–327, 2006.

[2] J. Miller, A. D. McLachlan, and A. Klug. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J.*, 4(6):1609–1614, Jun 1985.

[3] L. Hedstrom. Serine protease mechanism and specificity. *Chem. Rev.*, 102(12):4501–4524, Dec 2002.

[4] J. P. Nilmeier, D. A. Kirshner, S. E. Wong, and F. C. Lightstone. Rapid catalytic template searching as an enzyme function prediction procedure. *PLoS ONE*, 8(5):e62535, 2013.

[5] E. C. Meng, B. J. Polacco, and P. C. Babbitt. Superfamily active site templates. *Proteins*, 55(4):962–976, Jun 2004.

[6] J. Hou, S. R. Jun, C. Zhang, and S. H. Kim. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl. Acad. Sci. U.S.A.*, 102(10):3651–3656, Mar 2005.

[7] I. Jonassen, I. Eidhammer, and W. R. Taylor. Discovery of local packing motifs in protein structures. *Proteins*, 34(2):206–219, Feb 1999.

[8] C. Zhou, P. Meysman, B. Cule, K. Laukens, and B. Goethals. Discovery of Spatially Cohesive Itemsets in Three-Dimensional Protein Structures. *IEEE/ACM Trans Comput Biol Bioinform*, 11(5):814–825, 2014.

[9] P. Meysman, C. Zhou, B. Cule, B. Goethals, and K. Laukens. Mining the entire Protein DataBank for frequent spatially cohesive amino acid patterns. *BioData Min*, 8:4, 2015.