Thinking outside the informatics box: Computed chemical properties for protein function annotation

Caitlyn L. Mills, Lydia A. Ruffner, Penny J. Beuning and Mary Jo Ondrechen

Department of Chemistry and Chemical Biology, Northeastern University, Boston MA 02115 USA

There are now over 14,000 Structural Genomics (SG) protein structures deposited in the Protein Data Bank (PDB) and most of these are of unknown or uncertain biochemical function. Reliable computational methods for the prediction of the function of protein structures is an important current need. Typically, functions are assigned using informatics-based approaches. The annotation of protein function by automated means has led to high rates of misannotations in some databases [1]. Here we present a complementary and powerful approach **based on computed chemical properties of the individual residues** in a protein structure. Partial Order Optimum Likelihood (POOL) is used to predict the residues in the query protein structure that are important for catalysis. Typically these include the residues in the first layer that make direct contact with the substrate molecule(s) and also some residues in the second and third layers that play supporting roles in the catalytic process [2]. Then, for proteins of known biochemical function, Structurally Aligned Local Sites of Activity (SALSA) [3] places the POOL-predicted residues into local structural alignments to establish chemical signatures – local arrays of active residues that are common to proteins of the same function. The POOL-predicted residues of the query (SG) protein are then aligned with the local chemical signatures for the different functional types. These alignments, each SG protein against each functional family, are scored in order to predict the most likely function of the SG proteins. Results are reported for the SG members of the Ribulose Phosphate Binding Barrel (RPBB), Clp-Crotonase, and Haloacid Dehalogenase superfamilies. While we find the SG proteins in the RPBB superfamily to be well annotated, we predict very high annotation error rates (about 75%) in the Clp-Crotonase superfamily. Of particular interest are cases of predicted misannotation, where our prediction differs from that of the assigned function. Experimental testing of our predictions is performed by direct biochemical assays. Our annotations are shown to be correct for the cases that have been tested to date.

Acknowledgments: This work has the support of the National Science Foundation under grant number CHE-1305655, MathWorks, Inc. and a PhRMA Foundation Fellowship awarded to CLM.

References

[1] A.M. Schnoes, S.D. Brown, I. Dodevski, and P.C. Babbitt, *PLoS Comp Biol* **5**(12), e1000605 (2009).

[2] H.R. Brodkin, N.A. DeLateur, S. Somarowthu, C.L. Mills, W.R. Novak, P.J. Beuning, D. Ringe, and M.J. Ondrechen, *Protein Sci* **24**, 762-778 (2015).

[3] Z. Wang, P. Yin, J. Lee, R. Parasuram, S. Somarowthu, and M.J. Ondrechen, *BMC Bioinformatics* **14**(Suppl 3), S13 (2013).