COMPUTATIONAL APPROACHES TO UNDERSTANDING THE EVOLUTION OF MOLECULAR FUNCTION

Yana Bromberg

Department of Biochemistry and Microbiology, Rutgers University
New Brunswick, New Jersey, U.S.A.

Matthew W. Hahn[†], Predrag Radivojac

[†]*Department of Biology, Indiana University*
Department of Computer Science and Informatics, Indiana University
Bloomington, Indiana, U.S.A.

1. Introduction

Understanding the function of biological macromolecules and their interactions is a grand challenge of modern biology, and a key foundation for biomedical research.[1,2] It is now evident that the function of these molecules, in isolation or in groups, can be productively studied in the context of evolution.[3,4] Therefore, understanding how these molecules and their functions evolve is an important step in understanding the specific events that lead to observable changes in molecular and biological processes.

With the advent of high-throughput technologies and the rapid accumulation of molecular data over the past several decades, the evolution of molecular function can be systematically studied at multiple levels. This includes the evolution of protein structure, 3D organization and dynamics, protein and gene expression, as well as the higher-level organization of function contained within pathways.[5–11] New experiments using the latest gene-editing technologies (such as CRISPR-Cas9) have also made it possible to directly test hypotheses about function in almost any organism.[12] Combining these data with theory and computational tools taken from evolutionary biology and related fields has led to an explosion in the study of how function evolves.

2. Overview of Contributions

Our session includes four accepted papers covering a variety of the subjects in this field. The papers address biological questions from metabolic processes to the evolution of duplicated genes; they use computational methods ranging from learning functions on biological networks to the optimal way to choose clustering parameters to identify homologs. Bowerman et al. investigate a set of about one hundred fully sequenced bacterial species mapped onto a space of metabolic variants via a literature search. They subsequently use these data to learn metabolic signatures among these species, an approach that can ultimately lead to a predictive system of metabolic potential for any bacterial species. Cao and Cowen study protein function transfer within a single species and ask under what conditions it leads to accurate prediction. Several sequence, network, and evolutionary features were examined to conclude that the level of sequence divergence is the major determinant of accurate function transfer among within-species paralogs in yeast. The paper relates to several earlier studies addressing evolutionary relationships and functional similarity.[13–17] Wang et al. present and evaluate a new approach for protein function prediction. Their method is based on amino acid sequences and protein-protein interaction networks over multiple species, integrated into a single heterogeneous network. Network integration is often challenging to formalize considering practical problems such as missing data, sample selection bias, and noise in available protein-protein interactions. Nevertheless, the approach showed good performance upon data integration and provided the insight that the combination of data sources contributed to increased accuracy. Finally, Wiwie and Röttger study the behavior and performance of several clustering algorithms in the context of detecting protein families in similarity graphs. Protein clustering is difficult owing to the unequal sizes of homologous families and the sensitivity of clusters to the parameters of the algorithm. They show that the original data can, in principle, be used to predict clustering performance but also highlight difficulties in finding optimal clustering parameters.

References

1. R. Rentzsch and C. A. Orengo, *Trends Biotechnol* **27**, 210 (2009).
2. P. Radivojac *et al.*, *Nat Methods* **10**, 221 (2013).
3. J. A. Eisen, *Genome Res* **8**, 163 (1998).
4. M. Pellegrini *et al.*, *Proc Natl Acad Sci U S A* **96**, 4285 (1999).
5. N. V. Grishin, *J Struct Biol* **134**, 167 (2001).
6. E. V. Koonin, *Annu Rev Genet* **39**, 309 (2005).
7. D. A. Drummond *et al.*, *Proc Natl Acad Sci U S A* **102**, 14338 (2005).
8. C. Pal *et al.*, *Nat Rev Genet* **7**, 337 (2006).
9. M. E. Peterson *et al.*, *Protein Sci* **18**, 1306 (2009).
10. W. Qian *et al.*, *Proc Natl Acad Sci U S A* **108**, 8725 (2011).
11. C. Park *et al.*, *Proc Natl Acad Sci U S A* **110**, E678 (2013).
12. J. A. Doudna and E. Charpentier, *Science* **346**, p. 1258096 (2014).
13. S. Mika and B. Rost, *PLoS Comput Biol* **2**, p. e79 (2006).
14. R. A. Studer and M. Robinson-Rechavi, *Trends Genet* **25**, 210 (2009).
15. N. L. Nehrt *et al.*, *PLoS Comput Biol* **7**, p. e1002073 (2011).
16. A. M. Altenhoff *et al.*, *PLoS Comput Biol* **8**, p. e1002514 (2012).
17. G. Plata and D. Vitkup, *Nucleic Acids Res* **42**, 2405 (2014).