

Producing a Molecular Parts List for Systems Biology

Heidi J. Sofia, Ph.D.

Computational Biochemistry, Biophysics & Biology
Pacific Northwest National Laboratory
Richland, WA 99352

Heidi.Sofia@pnl.gov

Jorge F. Reyes-Spindola

Information Sciences & Engineering
Pacific Northwest National Laboratory
Richland, WA 99352

Jorge.Reyes.Spindola@pnl.gov

ABSTRACT

Completed genome sequences enable the definition of a Molecular Parts List describing the components and the cellular machinery for each organism. Assigning functions to gene and protein sequences is based on experimental data and computational features extracted largely from evolutionary relationships. Many systems biology methods such as global expression experiments with gene arrays or proteomics, and modeling and simulation efforts assume that a complete Molecular Parts List is in place for full interpretation of the results. However, the widely varying quality of protein function data as it currently exists unnecessarily limits the utility of these powerful global methods.

The current paradigm for protein function assignment and the distribution of this resource is very inefficient. Each scientist is faced with large-scale annotations that are a complex mix of correct, incorrect, misleading, missing, and out-of-date information. Typically, for any individual to evaluate the reliability of a given assignment, multiple bioinformatics analyses and text searches of the biomedical literature must be repeated, and these efforts may become dated as soon as they are completed.

We are developing a new approach to large-scale protein function assignment based on strategies in information integration and visualization. Functional assignments can be directly linked to an interactive “view” that summarizes the evidence supporting a conclusion in a way that a biologist can interpret. Automated agents can update the collection and display of data so that the

reliability of any particular assignment can be assessed directly at any time.

We describe a data-mining visualization applied to the new Radical SAM superfamily, currently at 854 members, which we discovered using iterative profile searches (1). These unusual proteins employ radical chemistry to perform diverse difficult reactions. They include well-characterized examples such as the “adenosyl radical” proteins lysine 2,3-aminomutase (LAM), biotin synthase (BioB), lipoic acid synthase (LipA), and the activating enzymes for pyruvate formate-lyase and the anaerobic ribonucleotide reductase. Also found are many partially characterized proteins such as the nitrogenase cofactor biosynthesis protein NifB and the heme biosynthesis protein HemN. At least half the proteins are unknown in function. We describe how new approaches to discovery based on information integration and visualization applied to distant sequence similarity relationships as well as other data types such as additional sequence-based patterns, expression profiles, biochemical and genetic features, and the biomedical literature can lead to enhanced problem-solving capabilities in functional assignment.

REFERENCES

- [1] Sofia, H.J., Chen, G., Hetzler, B.G., Reyes-Spindola, J.F. and Miller, N.E. Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods (2001) *Nucleic Acids Research* 29: 1097-1106.