# Discovery of Protein Functions and Interactions from Structures, Sequences and Text in Enzymes, Malaria and Cancer

Angela Wilkins[a], Scott Spangler[c], Andreas Martin Lisewski[a], Shivas Amin[b] and Olivier Lichtarge[a]

[a]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, [b]University of St Thomas, Houston, 77006 and [c]IBM Research Center, Almaden, CA 95120

We present four complementary studies that show the central role of network representations in molecular biology. Whether the information conveyed by edges that connect protein nodes represents (1) structural and evolutionary mimicry, (2) high throughput experimental data on interaction and homology, (3) contextual word similarities in papers, or (4) other biological relationships mined from text; network analysis using a diversity of algorithm can pass information from areas richly annotated to areas that are less so. These applications span enzymology, malaria and cancer, and all are supported by experimental discoveries guided by network predictions. Together these studies suggest that network views of biological processes are fundamental tools with the power to integrate high throughput experimental biology (BIG DATA) with the entire corpus of the biomedical literature (BIG LITERATURE) in order to guide discoveries through automated hypotheses generation.

## 1. Structure-based network discovery of gene function: Prediction And Validation of Enzyme Function and Substrate Specificity in Protein Structures through Structural Genomics Networks of Local Evolutionary Importance.

Structural Genomics aims to elucidate protein structures in order to identify their functions. Unfortunately, the variation of just a few residues can be enough to alter activity or binding specificity and limit the functional resolution of annotations based on sequence and structure; in enzymes, substrates are especially difficult to predict. Here, large-scale controls and direct experiments show that a protein network of local similarity of five or six residues selected because they are evolutionarily important, and on the protein surface, can suffice to identify an enzyme activity and substrate. Large-scale retrospective validation is confirmed by direct prospective experimental discovery: A motif of five residues predicted that a previously uncharacterized *Silicibacter sp* protein was a carboxylesterase for short fatty acyl chains, similar to hormone-sensitive lipase-like proteins that share less than 20% sequence identity. Assays and directed mutations confirmed this activity and showed that the motif was essential for catalysis and substrate specificity. We conclude that evolutionary and structural information may be combined in a Structural Genomics network to transfer information on molecular function across edges representing novel, shared evolutionary motifs of mixed catalytic and non-catalytic residues that identify enzyme activity and substrate specificity.

The significance of this work is that many proteins solved by Structural Genomics have low sequence identity to other proteins and cannot be assigned functions. Strikingly, this problem, computational approach creates structural motifs of a few evolutionarily important residues on the fly, and these motifs probe local geometric and evolutionary similarities in other protein structures to detect functional similarities. This approach does not require prior knowledge of functional mechanisms and is highly accurate in computational benchmarks when annotations rely on homologs with low sequence identity. Its accuracy is strongly supported by biochemical and mutagenesis studies to validate two predictions of unannotated proteins.

## 2. Sequence-based network discovery of gene function: Supergenomic network compression and the discovery of EXP1 as a glutathione transferase inhibited by artesunate

A central and complex problem in biology is to identify gene function. This can be guided by large supergenomic networks of interactions and ancestral relationships among genes. We show here that these biological networks are compressible. They can be shrunk dramatically, by eliminating redundant evolutionary relationships, and efficiently because their number of compressible cliques rises linearly with network size rather than exponentially as in other complex networks. Compression enables global network analysis to computationally harness hundreds of interconnected genomes and to produce novel functional predictions. As a demonstration, we show that the essential but functionally uncharacterized Plasmodium falciparum antigen EXP1 is a membrane glutathione S-transferase. EXP1 efficiently degrades cytotoxic hematin, is potently inhibited by artesunate with a half-maximal inhibitory concentration near 1 nM, and is associated with altered artesunate susceptibility in cultured malaria parasites. These data make EXP1 a possible molecular target to a frontline antimalarial drug.

## 3. Text-based Automated Hypothesis Generation Based on Mining Scientific Literature

Keeping up with the ever-expanding flow of data and publications is untenable and poses a fundamental bottleneck to scientific progress. Current search technologies typically find many relevant documents, but they do not extract and organize the information content of these documents or suggest new scientific hypotheses based on this organized content. We present an initial case study on KnIT, a prototype system that mines the information contained in the scientific literature, represents it explicitly in a queriable network, and then further reasons upon these data to generate novel and experimentally testable hypotheses. KnIT combines entity detection with neighbor-text feature analysis and with graph-based diffusion of information to identify potential new properties of entities that are strongly implied by existing relationships. We discuss a successful application of our approach that mines the published literature to identify new protein kinases that phosphorylate the protein tumor suppressor p53. Retrospective analysis demonstrates the accuracy of this approach and ongoing laboratory experiments suggest that kinases identified by our system may indeed phosphorylate p53. These results establish proof of principle for automated hypothesis generation and discovery based on text mining of the scientific

## 4. Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature

We present KnIT, the Knowledge Integration Toolkit, a system for accelerating scientific discovery and predicting previously unknown protein-protein interactions. Such predictions enrich biological research and are pertinent to drug discovery and the understanding of disease. Unlike a prior study, KnIT is now fully automated and demonstrably scalable. It extracts information from the scientific literature, automatically identifying direct and indirect references to protein interactions, which is knowledge that can be represented in network form. It then reasons over this network with techniques such as matrix factorization and graph diffusion to predict new, previously unknown interactions. The accuracy and scope of KnIT's knowledge extractions are validated using comparisons to structured, manually curated data sources as well as by performing retrospective studies that predict subsequent literature discoveries using literature available prior to a given date. The KnIT methodology is a step towards automated hypothesis generation from text, with potential application to other scientific domains.

## References

**1.** *Prediction and experimental validation of enzyme substrate specificity in protein structures* (2013) Shivas R Amin, Serkan Erdin, R Matthew Ward, Rhonald C Lua, Olivier Lichtarge. **PNAS** 110(45):E4195-E4202

**2.** *Supergenomic network compression and the discovery of EXP1 as a glutathione transferase inhibited by artesunate* (2014) Andreas Martin Lisewski, Joel P Quiros, Caroline L Ng, Anbu Karani Adikesavan, Kazutoyo Miura, Nagireddy Putluri, Richard T Eastman, Daniel Scanfeld, Sam J Regenbogen, Lindsey Altenhofen, Manuel Llinás, Arun Sreekumar, Carole Long, David A Fidock, Olivier Lichtarge. **Cell** (158)4:916-928.

**3**. *Automated hypothesis generation based on mining scientific literature* (2014). Scott Spangler, Angela D Wilkins, Benjamin J Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R Pickering, Austin Comer, Jeffrey N Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence Donehower, Ying Chen, Olivier Lichtarge. **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, NewYork (New York).**

**4**. *Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature* (2015). Meena Nagarajan, Angela D. Wilkins Benjamin J. Bachman, Ilya B. Novikov, Sheng Hua Bao, Peter J. Haas, María E. Terrón-Díaz, Sumit Bhatia, Anbu K. Adikesaven, Jacques J. Labrie, Sam Regenbogen, Christie M. Buchovecky, Curtis R. Pickering, Linda Kato, Andreas Martin Lisewski, Ana Lelescu, Houyin Zhang, Stephen Boyer, Griff Weber, Ying Chen, Lawrence Donehower, Scott Spangler, Olivier Lichtarge. **Proceedings of the 21th ACM SIGKDD international conference on Knowledge discovery and data mining, Sydney (Australia).**