INSTITUTE FOR SYSTEMS BIOLOGY

Final Scientific/ Technical Report

Award No. DE-FG02-07ER64327

Molecular Assemblies, Genes and Genomics Integrated Efficiently (MAGGIE): component 3: Systems Approach in a Multi-Organism Strategy..."

Nitin S. Baliga, Principal Investigator

5/26/2011

A final report including a detailed summary of accomplishments during the period of this grant, and publications resulting from research supported by the Department of Energy.

I. Executive Summary

We set ambitious goals to model the functions of individual organisms and their community from molecular to systems scale. These scientific goals are driving the development of sophisticated algorithms to analyze large amounts of experimental measurements made using high throughput technologies to explain and predict how the environment influences biological function at multiple scales and how the microbial systems in turn modify the environment. By experimentally evaluating predictions made using these models we will test the degree to which our quantitative multiscale understanding will help to rationally steer individual microbes and their communities towards specific tasks.

Towards this end we have made substantial progress towards understanding evolution of gene families, transcriptional structures, detailed structures of keystone molecular assemblies (proteins and complexes), protein interactions, biological networks, microbial interactions, and community structure.

Using comparative analysis we have tracked the evolutionary history of gene functions to understand how novel functions evolve. One level up, we have used proteomics data, high-resolution genome tiling microarrays, and 5' RNA sequencing to revise genome annotations, discover new genes including ncRNAs, and map dynamically changing operon structures of five model organisms: For *Desulfovibrio vulgaris Hildenborough*, *Pyrococcus furiosis*, *Sulfolobus solfataricus*, *Methanococcus maripaludis* and *Halobacterium salinarum NRC-1*. We have developed machine learning algorithms to accurately identify protein interactions at a near-zero false positive rate from noisy data generated using tagless complex purification, TAP purification, and analysis of membrane complexes. Combining other genome-scale datasets produced by ENIGMA (in particular, microarray data) and available from literature we have been able to achieve a true positive rate as high as 65% at almost zero false positives when applied to the manually curated training set. Applying this method to the data representing around a quarter of the fraction space for water soluble proteins in *D. vulgaris*, we obtained 854 reliable pair wise interactions. Further, we have developed algorithms to analyze and assign significance to protein interaction data from bait pull-down experiments and integrate these data with other systems biology data through associative biclustering in a parallel computing environment. We will "fill-in" missing information in these interaction data using a "Transitive Closure" algorithm and subsequently use "Between Commonality Decomposition" algorithm to discover complexes within these large graphs of protein interactions. To characterize the metabolic activities of proteins and their complexes we are developing algorithms to deconvolute pure mass spectra, estimate chemical formula for m/z values, and fit isotopic fine structure to metabolomics data. We have discovered that in comparison to isotopic pattern fitting methods restricting the chemical formula by these two dimensions actually facilitates unique solutions for chemical formula generators.

To understand how microbial functions are regulated we have developed complementary algorithms for reconstructing gene regulatory networks (GRNs). Whereas the network inference algorithms cMonkey and Inferelator developed enable de novo reconstruction of predictive models for GRNs from diverse systems biology data, the RegPrecise and RegPredict framework developed uses evolutionary comparisons of genomes from closely related organisms to reconstruct conserved regulons. We have integrated the two complementary algorithms to rapidly generate comprehensive models for gene regulation of understudied organisms. Our preliminary analyses of these reconstructed GRNs have revealed novel regulatory mechanisms and cis-regulatory motifs, as well as others that are conserved across species.

Finally, we are supporting scientific efforts in ENIGMA with data management solutions and by integrating all of the algorithms, software and data into a Knowledgebase. For instance, we have developed the RegPrecise database (http://regprecise.lbl.gov) which represents manually curated sets of regulons laying the basis for automatic annotation of regulatory interactions in closely related species. We are also in the midst of scaling up MicrobesOnline to handle the growing volume of sequence and functional genomics data. Over the last year our efforts have been focused on providing support for additional genomic and functional genomic data types. Similarly, we have developed several visualization tools to help with the exploration of complex systems biology datasets. A case in point is the Gaggle Genome Browser (GGB), which was enhanced with visualizations for plotting peptide detections and protein-DNA binding alongside transcriptome structure, plus the ability to interactively filter by signal intensity or p-value. Finally, we recognize that future advances to computational infrastructure cannot be anticipated and new software will be developed as dictated by scientific needs within ENIGMA and elsewhere. To account for this reality of how software environments evolve we have made advances to the Gaggle and Firegoose framework that enables interoperability and integration of diverse software and databases. Specifically, we have updated the R-goose package which provides connectivity between several Gaggle compliant bioinformatics tools and R; and prototyped a JSON based upgrade to the Gaggle protocol to make this environment extensible and more language neutral than the previous Java-based protocol.

II. Publications:

1. Yoon SH, Reiss DJ, Bare JC, Tenenbaum D, Pan M, Slagel J, Lim S, Hackett M, Menon AL, Adams MW, Barnebey A, Yannone SM, Leigh JA, and Baliga NS. "Parallel evolution of transcriptome architecture during genome reorganization" (Genome Research in review)
2. Robinson CK, Webb K, Kaur A, Jaruga P, Dizdaroglu M, Baliga NS, Place A, Diruggiero J. A Major Role for Non-Enzymatic Antioxidant Processes in the Radioresistance of Halobacterium salinarum. J Bacteriol. 2011 Jan 28.
3. Brooks AN, Turkarslan S, Beer KD, Yin Lo F, Baliga NS. Adaptation of cells to new environments. Wiley Interdiscip Rev Syst Biol Med. 2010 Dec 31

4. Tautenhahn R, Patti GJ, Kalisiak E, Miyamoto T, Schmidt M, Lo FY, McBee J, Baliga NS, Siuzdak G. metaXCMS: Second-Order Analysis of Untargeted Metabolomics Data. *Anal. Chem.*, Article ASAP DOI: 10.1021/ ac102980g Publication Date (Web): December 21, 2010
5. Schmid AK, Pan M, Sharma K, Baliga NS Two transcription factors are necessary for iron homeostasis in a salt-dwelling archaeon. Nucleic Acids Res: Nucleic Acids Research, 2010, 1–15 doi:10.1093/nar/gkq1211
6. Kaur A, Van PT, Busch CR, Robinson CK, Pan M, Pan WL, Reiss DJ, DiRuggiero J, Baliga NS. Coordination of frontline defense mechanisms under severe oxidative stress. Molecular Systems Biology 6:393.
7. Facciotti MT, Pang WL, Lo FY, Whitehead K, Koide T, Masumura K, Pan M, Amardeep Kaur, Larsen DJ, Reiss DJ, Hoang L, Kalisiak E, Northen T, Trauger SA, Siuzdak G, Baliga NS. Large scale physiological readjustment during growth enables rapid, comprehensive and inexpensive systems analysis. BMC Systems Biology, 2010 May 14;4:64
8. Kaur A, Van PT, Busch CR, Robinson CK, Pan M, Pang WL, Reiss DJ, DiRuggiero J, Baliga NS. Coordination of frontline defense mechanisms under severe oxidative stress. Mol Syst Biol. 2010 Jul;6:393.
9. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin AC.
10. Visualization of omics data for systems biology. Nat Methods. 2010 Mar;7(3 Suppl):S56-68.
11. Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo FY, Pratap A, Deutsch EW, Peterson A, Martin D, Baliga NS. Prevalence of transcription promoters within archaeal operons and coding sequences. Mol Syst Biol. 2009;5:285.
12. Whitehead K, Min P, Masumura K, Bonneau R, Baliga, NS. Diurnally entrained anticipatory behavior in Archaea. *PLoS One*. 2009;4(5):e5485. Epub 2009 Jun 19.
13. Schmid, AK, Pan, M, and Baliga, NS. Metabolic regulation
14. Koide T, Pang WL, Baliga NS. (2009), The role of predictive modelling in rationally re-engineering biological systems. *Nature Reviews Microbiology*. Apr;7(4):297-305.
15. Van P, Schmid AK, King N, Kaur A, Pan M, Whitehead K, Koide T, Facciotti M, Goo YA, Deutsch E, Reiss DJ, Mallick P, Baliga NS (2008) Halobacterium salinarium NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage. *Journal of Proteome Research*. Sep;7(9):3755-64
16. Baliga NS (2008) Perspectives: Microbiology - Scale of Prediction. *Science*. June 6: 320(5880): 1297-1298.
17. Reiss DJ, Facciotti MT, Baliga NS, (2008), Model-based deconvolution of genome-wide DNA binding. *Bioinformatics* 24, 396-403.
18. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson T, Shannon P, Johnson MH, Bare CJ, Longabaugh W, Vuthoori M, Whitehead K, Madar A, Suzuki L, Mori T, Chang DE, DiRuggiero J, Johnson CH, Hood L, Baliga NS, (2007), A predictive model for transcriptional control of physiology in a free living cell, *Cell* 131, 1354-65.
19. Bare JC, Shannon P, Schmid A, Baliga NS, (2007), The Firegoose: two-way integration of diverse data from different bioinformatics web resources with desktop applications. *BMC Bioinformatics*, 8, 456.
20. Schmid AK, Reiss DJ, Kaur A, Pan M, King N, Van PT, Hohmann L, Martin

DB, Baliga NS, (2007), The anatomy of microbial cell state transitions in response to oxygen. *Genome Research*. Oct; 17(10): 1399-413.
21. Facciotti MT, Reiss DJ, Pan M, Kaur A, Vuthoori M, Bonneau R, Shannon P, Srivastava A, Donohoe SM, Hood LE, Baliga NS, (2007), General transcription factor specified global gene regulation in archaea. *Proc Natl Acad Sci U S A.*, 104(11): 4630-4635.
22. Amy K. Schmid and Nitin S. Baliga (2007), Prokaryotic Systems Biology. *Cell Engineering* Vol. 5: Systems Biology. p395-423; M. Al-Rubeai and M. Fussenegger (ed.), Springer Press, New York, NY.
23. Kaur A, Pan M, Meislin M, Facciotti MT, El-Geweley R, Baliga NS. (2006), A systems view of haloarchaeal strategies to withstand stress from transition metals. *Genome Research*, 16:841-854. Reiss DJ, Baliga NS, Bonneau R. (2006), Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7:280.
24. Bonneau R, Reiss D, Shannon P, Facciotti MT, Hood L, Baliga NS, Thorsson V. (2006), The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data-sets. *Genome Biology*, 7:R36.
25. Shannon P, Reiss D, Bonneau R, Baliga N. (2006), Gaggle: An open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics*, 7:176.
26. Kenia Whitehead, Adrienne Kish, Min Pan, Amardeep Kaur, David J Reiss, Nichole King, Laura Hohmann, Jocelyne DiRuggiero, Nitin S Baliga (2006), An integrated systems approach for understanding cellular responses to gamma radiation. *Molecular Systems Biology*, 2:47.

There were no published conference papers or other public release of results.

III. Products Developed

The website links that reflect the results for the MAGGIE project are:

http://gaggle.systemsbiology.net/docs/

http://baliga.systemsbiology.net/drupal/content/enigma

http://baliga.systemsbiology.net/drupal/content/h2-regulation

The project did generate several network references:
a) C-monkey
b) Inferelator
c) Gaggle
d) Firegoose
e) Genome Browser

VI. Computer Modeling

Transcriptional regulatory network model, environment and gene regulatory inference network (EGRIN) version 1.0, cMonkey inferelator – intended use to predict response to new environments.

The performance criteria for the model related to the intended use was to test prediction on new experimental data.

The test results that demonstrated the models performance was met can be confirmed in the paper: Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson T, Shannon P, Johnson MH, Bare CJ, Longabaugh W, Vuthoori M, Whitehead K, Madar A, Suzuki L, Mori T, Chang DE, DiRuggiero J, Johnson CH, Hood L, Baliga NS, (2007), A predictive model for transcriptional control of physiology in a free living cell, *Cell* 131, 1354-65. We used Generalized Linear Model Regression. The theory behind the models is that biological networks are modular. By using a Regression model we can discover environmental and genetic control of these models. The resulting documentation results are open source that can be seen by viewing: http://baliga.systemsbiology.net

V. Training

CURRENT GRADUATE STUDENTS AND POSTDOCS

Aaron Brooks	GS. University of Washington, Seattle, WA
Karlyn Beer	GS. University of Washington, Seattle, WA
Fang-Yin Lo	GS. University of Washington, Seattle, WA
Sung Ho Yoon	Postdoc. Ph.D. (2002) KAIST, Seoul, S. Korea
Serdar Turkarslan	Postdoc. Ph.D. (2007) UPenn, Philadelphia, PA
Danielle Miller	Postdoc. Ph.D. (2008) UT Southwestern, Dallas, TX
Lee Pang	Postdoc. Ph.D. (2007) UCSD, San Diego, CA
Chris Plaisier	Postdoc. Ph.D. (2009) UCLA, Los Angeles, CA
Elisabeth Wurtmann	Postdoc. Ph.D. (2010) Yale University, New Haven, CT
Justin Ashworth	Postdoc. Ph.D. (2010) UW, Seattle, WA

FORMER LAB MEMBERS

Richard Bonneau	Assistant Professor, New York University, NY, NY
Amy Schmid	Assistant Professor, Duke University, Durham, NC
Tie Koide	Assistant Professor, Universidad de Sao Paolo, Sao Paolo, Brazil
Marc Facciotti	Assistant Professor, UC Davis, Davis, CA