COMPUTATIONAL ANALYSIS AND SIMULATION OF BACTERIAL MOLECULAR NETWORKS

1. DOE award: FG02-01ER25500

PI: Andrey Rzhetsky (Columbia University)

Co-PI: Dimitris Anastassiou (Columbia University)

The project covered interval between 09/29/2001 and 08/28/2005.

2. No distribution limitations specified

3. The first sub-project resulted in an original probabilistic model allowing assignment of a likelihood value to an arbitrary molecular-interaction network for a set of proteins that had known amino-acid sequences. The model was polished and improved over a series of published studies (e.g., Iossifov et al, 2004; Gomez at al. 2002). The main assumptions of this set of models is as follows: Information on molecular networks, such as networks of interacting proteins, comes from diverse sources that contain remarkable differences in distribution and quantity of errors. We developed a probabilistic model useful for predicting protein interactions from heterogeneous data sources. The model describes stochastic generation of protein-protein interaction networks with real-world properties, as well as generation of two heterogeneous sources of protein-interaction information: research results automatically extracted from the literature and yeast two-hybrid experiments. Based on the domain composition of proteins, we used the model to predict protein interactions for pairs of proteins for which no experimental data are available. We further explored the prediction limits, given experimental data that cover only part of the underlying protein networks. This approach can be extended naturally to include other types of biological data sources.

The second sub-project (see Cavelier & Anastassiou, 2004, 2005) started with assumption that finding the causality and strength of connectivity in transcriptional regulatory networks from time-series data will provide a powerful tool for the analysis of cellular states. We developed the design of tools for the evaluation of the network's model structure and parameters. The most effective tools are found to be based on evolution strategies. We evaluated models of increasing complexity, from lumped, algebraic phenomenological models to Hill functions and thermodynamically derived functions. These last functions provide the free energies of binding of transcription factors to their operators, as well as cooperativity energies. Optimization results were based on published experimental data from a synthetic network in Escherichia coli are presented. The free energies of binding and cooperativity found by our tools are in the same physiological ranges as those experimentally derived in the bacteriophage lambda system. We also use time-series data from high-density oligonucleotide

microarrays of yeast meiotic expression patterns. The algorithm appropriately finds the parameters of pairs of regulated regulatory yeast genes, showing that for related genes an overall reasonable computation effort is sufficient to find the strength and causality of the connectivity of large numbers of them.

The third subproject was associated with analysis and processing of text-mined data: The immense growth in the volume of research literature and experimental data in the field of molecular biology calls for efficient automatic methods to capture and store information. In recent years, several groups have worked on specific problems in this area, such as automated selection of articles pertinent to molecular biology, or automated extraction of information using natural-language processing, information visualization, and generation of specialized knowledge bases for molecular biology. GeneWays is an integrated system that combines several such subtasks. It analyzes interactions between molecular substances, drawing on multiple sources of information to infer a consensus view of molecular networks. GeneWays is designed as an open platform, allowing researchers to query, review, and critique stored information (Rzhetsky et al. 2004).

These studies add to our current understanding of complex molecular networks, their dynamic properties, and uncertainty associated with individual statements.

4. The aims of the project were met in full.

5. The activities regarding hypotheses, methods, and tests associated with our work are documented in publications listed in 6a.

6a. The following is the list of publications that stemmed from the DOE award.

[1] Rzhetsky, A. and W. M. Fitch, Listening to viral tongues: comparing viral trees using a stochastic context-free grammar, *Mol Biol Evol* **22** (2005), pp. 905-913.

[2] Cokol, M., I. Iossifov, C. Weinreb and A. Rzhetsky, Emergent behavior of growing knowledge about molecular interactions, *Nat Biotechnol* **23** (2005), pp. 1243-1247.

[3] Rzhetsky, A., I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. A. Duboue, W. Weng, W. J. Wilbur, V. Hatzivassiloglou and C. Friedman, GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data, *J Biomed Inform* **37** (2004), pp. 43-53.

[4] Krauthammer, M., C. A. Kaufmann, T. C. Gilliam and A. Rzhetsky, Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease, *Proc Natl Acad Sci U S A* **101** (2004), pp. 15148-15153.

[5] Iossifov, I., M. Krauthammer, C. Friedman, V. Hatzivassiloglou, J. S. Bader, K. P. White and A. Rzhetsky, Probabilistic inference of molecular networks from noisy data sources, *Bioinformatics* **20** (2004), pp. 1205-1213.

[6] Chien, M., I. Morozova, S. Shi, H. Sheng, J. Chen, S. M. Gomez, G. Asamani, K. Hill, J. Nuara, M. Feder, J. Rineer, J. J. Greenberg, V. Steshenko, S. H. Park, B. Zhao, E. Teplitskaya, J. R. Edwards, S. Pampou, A. Georghiou, I. C. Chou, W. Iannuccilli, M. E. Ulz, D. H. Kim, A. Geringer-Sameth, C. Goldsberry, P. Morozov, S. G. Fischer, G. Segal, X. Qu, A. Rzhetsky, P. Zhang, E. Cayanis, P. J. De Jong, J. Ju, S. Kalachikov, H. A. Shuman and J. J. Russo, The genomic sequence of the accidental pathogen Legionella pneumophila, *Science* **305** (2004), pp. 1966-1968.

[7] Roth, C. W., I. Holm, M. Graille, P. Dehoux, A. Rzhetsky, P. Wincker, J. Weissenbach and P. T. Brey, Identification of the Anopheles gambiae ATP-binding cassette transporter superfamily genes, *Mol Cells* **15** (2003), pp. 150-158.

[8] Karev, G. P., Y. I. Wolf, A. Rzhetsky, F. S. Berezovskaya and E. V. Koonin, Mathematical Modeling of the Evolution of Domain Composition of Proteomes: A Birth-and-Death Process with Innovation In: M. Y. Galperin and E. V. Koonin, Editors, *Frontiers in Computational Genomics* (2003).

[9] Gomez, S. M., W. S. Noble and A. Rzhetsky, Learning to predict protein-protein interactions from protein sequences, *Bioinformatics* **19** (2003), pp. 1875-1881.

[10] Yu, H., V. Hatzivassiloglou, A. Rzhetsky and W. J. Wilbur, Automatically identifying gene/protein terms in MEDLINE abstracts, *J Biomed Inform* **35** (2002), pp. 322-330.

[11] Yu, H., V. Hatzivassiloglou, C. Friedman, A. Rzhetsky and W. J. Wilbur, Automatic extraction of gene and protein synonyms from MEDLINE and journal articles, *Proc AMIA Symp* (2002), pp. 919-923.

[12] Krauthammer, M., P. Kra, I. Iossifov, S. M. Gomez, G. Hripcsak, V. Hatzivassiloglou, C. Friedman and A. Rzhetsky, Of truth and pathways: chasing bits of information through myriads of articles, *Bioinformatics* **18 Suppl 1** (2002), pp. S249-S257.

[13] Karev, G. P., Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya and E. V. Koonin, Birth and death of protein domains: A simple model of evolution explains power law behavior, *BMC Evol Biol* **2** (2002), p. 18.

[14] Gomez, S. M. and A. Rzhetsky, Towards prediction of complete protein-protein interaction networks, *Pac Symp Biocomput* (2002), pp. 413-424.

[15] Friedman, C., P. Kra and A. Rzhetsky, Two biomedical sublanguages: a description based on the theories of Zellig Harris, *J Biomed Inform* **35** (2002), pp. 222-235.

[16] Christophides, G. K., E. Zdobnov, C. Barillas-Mury, E. Birney, S. Blandin, C. Blass, P. T. Brey, F. H. Collins, A. Danielli, G. Dimopoulos, C. Hetru, N. T. Hoa, J. A. Hoffmann, S. M. Kanzok, I. Letunic, E. A. Levashina, T. G. Loukeris, G. Lycett, S. Meister, K. Michel, L. F. Moita, H. M. Muller, M. A. Osta, S. M. Paskewitz, J. M. Reichhart, A. Rzhetsky, L. Troxler, K. D. Vernick, D. Vlachou, J. Volz, C. Von Mering, J. Xu, L. Zheng, P. Bork and F. C. Kafatos, Immunity-related genes and gene families in Anopheles gambiae, *Science* **298** (2002), pp. 159-165.

[17] Tammur, J., C. Prades, I. Arnould, A. Rzhetsky, A. Hutchinson, M. Adachi, J. D. Schuetz, K. J. Swoboda, L. J. Ptacek, M. Rosier, M. Dean and R. Allikmets, Two new genes from the human ATP-binding cassette transporter superfamily, ABCC11 and ABCC12, tandemly duplicated on chromosome 16q12, *Gene* **273** (2001), pp. 89-96.

[18] Scott, K., R. Brady, A. Cravchik, P. Morozov, A. Rzhetsky, C. Zuker and R. Axel, A chemosensory gene family encoding candidate gustatory and olfactory receptors in Drosophila, *Cell* **104** (2001), pp. 661-673.

[19] Rzhetsky, A. and P. Morozov, Markov chain Monte Carlo computation of confidence intervals for substitution-rate variation in proteins, *Pacif Symp Biocomp* **6** (2001), pp. 203-214.

[20] Rzhetsky, A. and S. M. Gomez, Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome, *Bioinformatics* **17** (2001), pp. 988-996.

[21] Hatzivassiloglou, V., P. A. Duboue and A. Rzhetsky, Disambiguating proteins, genes, and RNA in text: a machine learning approach, *Bioinformatics* **17 Suppl 1** (2001), pp. S97-S106.

[22] Gomez, S. M., S. H. Lo and A. Rzhetsky, Probabilistic prediction of unknown metabolic and signal-transduction networks, *Genetics* **159** (2001), pp. 1291-1298.

[23] Friedman, C., P. Kra, H. Yu, M. Krauthammer and A. Rzhetsky, GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, *Bioinformatics* **17 Suppl 1** (2001), pp. S74-S82.

[24] Dean, M., A. Rzhetsky and R. Allikmets, The human ATP-binding cassette (ABC) transporter superfamily, *Genome Res* **11** (2001), pp. 1156-1166.

[25] Annilo, T., J. Tammur, A. Hutchinson, A. Rzhetsky, M. Dean and R. Allikmets, Human and mouse orthologs of a new ATP-binding cassette gene, ABCG4, *Cytogenet Cell Genet* **94** (2001), pp. 196-201.

[26] Cavelier, G. & Anastassiou, D. Phenotype analysis using network motifs derived from changes in regulatory network dynamics. *Proteins* **60**, 525-546, doi:10.1002/prot.20538 (2005).

[27] Cavelier, G. & Anastassiou, D. Data-based model and parameter evaluation in dynamic transcriptional regulatory networks. *Proteins* **55**, 339-350, doi:10.1002/prot.20056 (2004).

[28] Anastassiou, D. Frequency-domain analysis of biomolecular sequences. *Bioinformatics (Oxford, England)* **16**, 1073-1081 (2000).

6b: N/A

6c: We fostered collaboration with Professor Dimitris Anastassiou's group (Department of Electrical Engineering, Columbia University).

6d N/A

6e N/A

6f N/A

7. We developed and implemented mathematical models (differential equation-based) for analysis of transcriptional regulatory networks (e.g., see Cavelier & Anastassiou, 2004, 2005), and Bayesian data integration (e.g., Iossifov et al 2004). The models were implemented in MatLab, and tested through cross-validation (probabilistic modeling) or comparison with real data (dynamics modeling).