The Quest for the Mechanisms of Life

Maria I. Klapa[1] and John Quackenbush[2,3]

[1] Department of Chemical Engineering, University of Maryland, College Park, MD 20742

[2] The Institute for Genomic Research, Rockville, MD 20850

[3] Department of Biochemistry, The George Washington University, Washington DC 20052

Submitted to Biotechnology and Bioengineering for the special issue on Systems Biology and Bioinformatics

ABSTRACT

The genomic revolution, manifested by the sequencing of the complete genome of many organisms, along with technological advances, such as DNA microarrays and developments in the analysis of proteins, metabolites and isotopic distribution patterns, challenged the conventional ways in which questions are approached in the biological sciences: (a) rather than examining a small number of genes and/or reactions at any one time, we can now analyze gene expression and protein activity in the context of systems of interacting genes and gene products, (b) comprehensive analysis of biological systems requires the integration of all cellular fingerprints: genome sequence, maps of gene expression, total protein production, metabolic output, and *in vivo* enzymatic activity, and (c) collecting, managing, and analyzing comparable data from various cellular profiles requires expertise from several fields that transcend traditional discipline boundaries. While researchers in systems biology have still to address difficult challenges in both experimental and computational arenas, they possess for the first time the opportunity to unravel the mechanisms of life. The enormous impact of these discoveries in diverse areas, such as metabolic engineering, strain selection, drug screening and development, bioprocess development, disease prognosis and diagnosis, gene and other medical therapies, is an obvious motivation for pursuing integrated analyses of cellular systems.

System Biology and The Genomic Revolution

Biology has evolved rapidly during the past fifty years, driven largely by advances in molecular biology coupled with developments in computer science. This convergence produced the genome revolution, allowing us to determine the complete genome sequence, and through it complete gene catalogues, for a number of important organisms, including humans. Genome sequencing along with other advances, such as the development of DNA microarrays [1-3], which allow the simultaneous measurement of the expression of every single gene in a cellular genome, and mass spectral analysis of proteins [4-8], metabolites [9-12] and isotopic tracer distribution patterns [13-15], have challenged the conventional paradigm of biological research. Rather than examining a small number of genes and/or reactions at any one time, we can now begin to look at gene expression and protein activity in the context of networks and systems of interacting genes and gene products. Because our knowledge of this domain is still largely rudimentary, investigations are now routinely moving from being "hypothesis-driven" to being "data-driven" with analysis based on a search for biologically relevant patterns. These technological advances have created enormous opportunities for accelerating the pace of science. One can now envision the possibility of obtaining a comprehensive picture of the mechanisms underlying the cellular function, its regulation, and the interactions of an organism with its environment.

While the greatest attention has been paid to gene sequence and transcriptional expression analysis using microarrays, it is becoming increasingly clear that these alone cannot be used to accurately determine cellular function. Rather, a comprehensive analysis of biological systems requires the integration of all fingerprints of cellular

function: genome sequence, maps of gene expression, total protein production, metabolic output, and *in vivo* enzymatic expression (activity). While each of these has significant value on its own, the picture that emerges from any single approach is quite limited in nature. Gene transcription is a necessary but not sufficient condition for high *in vivo* protein production. Regulation of translation, RNA and protein stability, and post-translational modifications can alter the linear relationship between message and the corresponding protein [16-18]. Additionally, a protein could be present in high concentration, but it may lack the requisite conditions (substrate concentration, cofactors, etc.) for activity in the actual cellular environment [19-20]. Integration of all of these profiles for a systematically perturbed cellular system can provide insight about the function of unknown genes, the relationship between gene and metabolic regulation and even the reconstruction of the gene regulation network [21].

Holistic analyses of biological systems, however, require a change in the way in which questions are approached in the biological sciences. Collecting, managing, and analyzing comparable data from various cellular profiles requires expertise from several fields that transcend traditional discipline boundaries, including engineering and computer science, statistics and applied mathematics, and chemistry, physics, and biology. This "systems biology" approach will be the framework for the training of a new generation of researchers in the life sciences who will be able to work, interact and collaborate in a very diverse and highly interdisciplinary environment.

Post-Genomic Research – Challenges, Opportunities, Directions

Despite the importance of integrated genomic, proteomic and metabolic studies, very few experiments have been done to date that actually combine information from

multiple cellular profiles. Most recent work has focused on one analysis of a single data type, or at best a combination of genomic and proteomic profiles. One of the main reasons is that we presently lack both the conceptual understanding and the computational tools that would allow the identification of cause-effect relationships between the gene and protein expression and phenotypic profiles. The development, however, of algorithms to address these questions cannot be accomplished in the absence of experimental data that monitor the cellular physiology under a variety of conditions at all stages of growth and levels of cellular function. Taking into consideration the different time-scales of the various biological processes, it is therefore very important to carefully design experiments that can provide comparable gene expression, protein production and metabolic function data that can lead to useful results. This will be closely tied to technological developments aiming at increasing and improving the experimental techniques and methodologies for the quantitative measurement of the cellular physiological state at each level of cellular function.

Even though DNA microarrays have revolutionized biological research, the measurement of gene expression profiles based on them remains a semi-quantitative process, which does not yet allow for absolute levels of gene expression to be identified. At the proteomic level, researchers are still working on the development of techniques for high throughput protein expression analysis, such as protein microarrays [22-23], which will provide a level of throughput similar to those obtained for transcriptional profiling with DNA arrays. Further, advances in protein crystallography will enable the high-throughput determination of the three-dimensional structure of

proteins [24] and significantly increase the quantity and quality of data in the public protein databases.

At the metabolic level, researchers are still in the search of techniques that might provide an enzymatic activity profile equivalent to those we can now obtain for gene expression and protein production. As it is highly unlikely that we will ever be able to develop an *in vivo* enzymatic activity chip, mapping the flux distribution through a metabolic reaction network [25] is the closest phenotypic equivalent to the type of data we can measure from available techniques for gene and protein expression. Fluxes are determined indirectly from the measurement of net excretion rates of extracellular metabolites and/or the use of isotopically labeled substrates [25]. All, however, comprehensive methods for the analysis of complex metabolic flux networks are presently primarily based on steady-state or pseudo steady-state assumptions in lack of accurate and extensive quantitative measurements of the intracellular metabolite concentrations and their isotopic tracer distribution. Advances in metabolic profiling [9-11], defined as the qualitative and quantitative detection (by Nuclear Magnetic Resonance Spectroscopy and Mass Spectrometry) of low molecular weight metabolites from the breakdown of the cellular macromolecules, are expected to enhance our understanding of metabolic activity under transient conditions [26-27]. This will lead to an increased number of integrated genomic and metabolic studies, which have been currently limited from to the lack of flux analysis methodologies for transient physiological conditions. Furthermore, technological and computational developments for metabolic characterization at the micro-scale [28] will increase dramatically the

number and type of examined physiological conditions opening enormous opportunities in the area of comparative biological studies.

Efficient use of the big load of data generated from systems biology studies will require development of extended databases that can effectively capture and integrate genomic, proteomic and phenotypic data. Currently there exist databases that store DNA and protein sequence data, protein three-dimensional structure, and metabolic pathway structure and stoichiometry, but it is still extremely difficult to link information across these diverse resources. Furthermore, these databases should be expanded to accommodate gene and protein expression along with *in vivo* metabolic activity data representing many different physiological conditions. The analysis of biological systems and the development of theoretical models that describe and predict cellular function must be based on integrated data from a large number of experiments. As the microarray community has come to realize, this will require the development of standards for describing experimental conditions and for submitting data to public databases (see MIAME protocol [29]). A similar initiative is imperative for the accurate collection of large quantities of systems biology data as the hope is that this can lead to conclusions about the interrelationships of the various cellular functions that manifest themselves under the experimental conditions under study.

Further, there is a clear need for development of data visualization and mining software that can be used with diverse data types to explore the relationships that exist and to infer the presence of metabolic pathways. Such a system would integrate gene annotation and a variety of expression data to allow visualization of metabolic pathway activity at the transcriptional level, connecting each gene to the reactions that are

catalyzed by the enzyme it encodes. If one assumes a direct correlation between changes in gene expression and associated enzymatic activity as reflected by metabolic output (an assumption in obvious need of verification), gene expression data should allow the formulation of a tentative metabolic network to be further confirmed by additional metabolic activity studies, including assessment of *in vivo* metabolic pathway activity as it is measured in terms of fluxes or metabolite concentrations. Any observed inconsistencies, such as high levels of gene expression without a corresponding change in metabolic activity, or the converse, will provide powerful leads to assist in developing verified causal relationships of consequence to overall cell behavior. With such an approach, the first obvious application of combined profiling of metabolic activity and gene expression will be in tracing the origin of easily observed physiological changes, focusing on well understood metabolic pathways as a means of justifying this approach. Although such pathways are often considered well-known, they were derived when information about only a relatively small number of genes was available, and we anticipate that integrated whole genome analyses will overturn many of the widely held assumptions about genetic and metabolic interrelationships.

Finally, we believe that at this stage, when integrated analyses are still in their infancy, appropriate model biological systems should be selected and used to validate software modules and computational algorithms developed from the combination of data. Short, well-controlled pathways, relatively isolated from the rest of metabolism or those well-studied with respect to their genomic and metabolic regulation should be used as test models. Experiments should be conducted in such a way to assure that the observed changes in the physiological profiles of the cells are due only to the applied

perturbations and not to other variables. In addition it is only through the comparison of the predictions of a computational algorithm with expected data based on previous biological knowledge that the conclusions of such algorithms can be validated. Candidates for such analyses include portions of central carbon metabolism and amino acid biosynthesis. While these are not isolated parts of the cellular network, they are among the best studied, particularly in bacteria, where the genes associated with these reactions are usually the first to be annotated in sequenced genomes.

In conclusion, it is clear that the combination of gene expression, protein production and *in vivo* metabolic activity data, along with new, powerful experimental and computational analytical methodologies will provide unprecedented insight into the structure of the language which is used by the cell to communicate changes in the cellular environment to gene expression and vice versa. While researchers in systems biology have still to overcome many obstacles and address difficult challenges in both experimental and computational arenas, they possess for the first time the opportunity to unravel the mechanisms of life. The enormous impact of these discoveries and the smaller ones along the way in diverse areas, such as metabolic engineering, strain selection, drug screening and development, bioprocess development, disease prognosis and diagnosis, gene and other medical therapies, is an obvious motivation for pursuing integrated analyses of cellular systems using combinations of methods that provide insight into physiological profiles.

References

1. Fodor, S.P.A. 1997. Massively parallel genomics. *Science* **277**: 393-395.

2. Schena, M., Shalon, D., Heller, R., Chai, A., Brown P.O. 1996. Parallel human genome analysis: Microarray based expression monitoring of 1000 genes. *Proc. Nat'l Acad. Of Sciences, USA* **93**: 10614-10619.
3. Brown, P.O., Botstsein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**: 33-37.
4. Lopez F., Pichereaux C., Burlet-Schiltz O., Pradayrol L., Monsarrat B., Esteve J.P. 2003. Improved sensitivity of biomolecular interaction analysis mass spectrometry for the identification of interacting molecules. *Proteomics* **3**:402-12.
5. Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J. 2003. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol*. [epub ahead of print]
6. Manabe T. 2003. Analysis of complex protein-polypeptide systems for proteomic studies. *J Chromatogr B Analyt Technol Biomed Life Sci.* **787**:29-41.
7. Wang H., Hanash S. 2003. Multi-dimensional liquid phase based separations in proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci*. **787**:11-8.
8. Jones JJ, Stump MJ, Fleming RC, Lay JO Jr, Wilkins CL. 2003. Investigation of MALDI-TOF and FT-MS techniques for analysis of Escherichia coli whole cells. *Anal Chem*. **75**:1340-7.
9. Roessner U., Wagner C., Kopka J., Trethewey R., Willmitzer L. 2000. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* 23:131-142.

10. Fiehn O., Kopka J., Dormann P., Altmann T., Trethewey R.N., Wilmitzer L. 2000. Metabolite profiling for plant functional genomics. *Nature Biotech*. 18:1157-

11. Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie A. 2001. Metabolic Profiling Allows Comprehensive Phenotyping of Genetically or Environmentally Modified Plant Systems. *Plant Cell* 13:11-29

12. Taylor J, King RD, Altmann T, Fiehn O. 2002. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics* Suppl 2:S241-8

13. Christensen B, Nielsen J. (1999) Isotopomer Analysis Using GC-MS. Metab. Eng. 1, 282 (E8)-290 (E16)

14. Fischer E., Sauer U. 2003. Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. Eur. J. Biochem. 270, 880-891

15. Wittmann C, Heinzle E. (2001) Application of MALDI-TOF MS to lysine-producing *Corynebacterium glutamicum* - A novel approach for metabolic flux analysis. Eur. J. Biochem. 268, 2441-2455.

16. Serikawa KA, Xu XL, MacKay VL, Law GL, Zong Q, Zhao LP, Bumgarrner R., Morris DR. 2003. The transcriptome and its translation during recovery from cell-cycle arrest in *Saccharomyces cerevisiae*. *Mol Cell Proteomics*. [epub ahead of print]

17. Rossignol F, Solares M, Balanza E, Coudert J, Clottes E. 2003. Expression of lactate dehydrogenase A and B genes in different tissues of rats adapted to chronic hypobaric hypoxia. *J Cell Biochem*. **89**:67-79.

18. Rehfeld JF, Goetze JP. 2003. The posttranslational phase of gene expression: new possibilities in molecular diagnosis. *Curr Mol Med*. **3**:25-38

19. Fell D. 1997. Understanding the control of metabolism. Portland Press Ltd, London
20. Stephanopoulos G., Aristidou A., Nielsen J. 1998. Metabolic Engineering. Academic Press, San Diego
21. Ideker T., Thorsson V., Ranish J.A., Christmas R., Buhler J., Eng J.K., Bumgarner R., Goodlett D.R., Aebersold R., Hood L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292: 929-934
22. Koopmann JO, Blackburn J. High affinity capture surface for matrix-assisted laser desorption/ionisation compatible protein microarrays. 2003. *Rapid Commun Mass Spectrom* **17**:455-62
23. Lal SP, Christopherson RI, dos Remedios CG. 2002. Antibody arrays: an embryonic but rapidly growing technology. *Drug Discov Today* **7(18 Suppl)**: S143-9
24. Yee A, Pardee K, Christendat D, Savchenko A, Edwards AM, Arrowsmith CH. 2003. Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. *Acc Chem Res.* **36**:183-9.
25. Stephanopoulos, Gregory. 1998. Metabolic Fluxes and Metabolic Engineering. Metabolic Engineering 1: 1-10
26. Fiehn O, Weckwerth W. 2003. Deciphering metabolic networks. *Eur J Biochem*. **270**: 579-88.
27. Hans MA, Heinzle E, Wittmann C. 2003. Free intracellular amino acid pools during autonomous oscillations in *Saccharomyces cerevisiae*. *Biotechnol Bioeng*. **82**:143-51

28. John GT, Klimant I, Wittmann C, Heinzle E. 2003. Integrated optical sensing of dissolved oxygen in microtiter plates: A novel tool for microbial cultivation. *Biotechnol Bioeng.* **81**: 829-36.

29. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet.* **29**:365-71.