IDENTIFICATION OF ABERRANT PATHWAY AND NETWORK ACTIVITY FROM HIGH-THROUGHPUT DATA

M. F. OCHS

Departments of Oncology and Health Science Informatics, Johns Hopkins University, Baltimore, MD 19075, USA
**E-mail: mfo@jhu.edu*

R. KARCHIN

Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA
E-mail: karchin@jhu.edu

H. RESSOM

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Washington, DC 20057, USA
E-mail: hwr@georgetown.edu

R. GENTLEMAN

Bioinformatics and Computational Biology, Genentech South San Francisco, CA 94080, USA
E-mail: gentleman.robert@gene.com

The workshop focused on approaches to deduce changes in biological activity in cellular pathways and networks that drive phenotype from high-throughput data. Work in cancer has demonstrated conclusively that cancer etiology is driven not by single gene mutation or expression change, but by coordinated changes in multiple signaling pathways. These pathway changes involve different genes in different individuals, leading to the failure of gene-focused analysis to identify the full range of mutations or expression changes driving cancer development. There is also evidence that metabolic pathways rather than individual genes play the critical role in a number of metabolic diseases. Tools to look at pathways and networks are needed to improve our understanding of disease and to improve our ability to target therapeutics at appropriate points in these pathways.

Keywords: Signal pathways, metabolic pathways, disease, statistics

1. Introduction

Many complex databases are being developed and maintained to house genetic, epigenetic, genomic, and functional genomic data. Centralized resources such as the National Center for Biomedical Informatics (NCBI) are developing databases to integrate reads from next generation sequencing experiments, tumor-derived somatic DNA sequence variation, and single nucleotide polymorphisms (SNPs) or haplotypes significantly associated with disease phenotypes. Functional genomic data and methylation array data are being captured in the Gene Expression Omnibus (GEO) and ArrayExpress data repositories. The cancer genome atlas (TCGA) combines all these types of data together with detailed information about clinical

phenotypes. A vast amount of open-access data now allows data analysts and informaticists the opportunity to develop tools and perform initial demonstrations of their validity independently of new bench experiments. These resources now provide a unique opportunity for the development of tools suitable for analyzing data arising from complex biology.

A key focus in this workshop was the emergence of model-based analysis for high-throughput data. As an example of previous work, Chinnaiyan's group utilized prior knowledge on gene expression and TF binding in prostate cancer to identify a change in a key metabolite associated with prostate cancer progression.[1] Sarcosine was one of many metabolites to show substantial changes in levels during prostate cancer progression, however it is produced by GNMT, a methyl transferase with an androgen receptor binding site upstream. As androgen is known to play an important role in prostate cancer aggressiveness, this allowed prediction that sarcosine might serve as a marker of aggressiveness and potentially even be a driver of such aggressiveness, which was validated in cell line studies. The interactions modeled between the molecular components in this work relied on building a simple mechanistic model of the underlying biology, without which the discovery could not have been made. The focus in this workshop was on efforts to integrate data and build models on a much larger scale.

A particularly promising point of integration is the role of pathways in disease. Biological pathways provide a natural approach to the integration of multiple omics data as well as a means to identify the mechanism through which the effects of mutations, epigenetic variation, protein isoforms, and metabolic changes occur.

2. Pathways in Human Disease

Recognition that biological pathways are critical to understanding human disease emerged along with the elucidation of metabolic and cell signaling pathways by molecular biologists and biochemists. For example, the discovery of the role of MAPK kinases in response to external signals[2] and the later elucidation of the proliferation response due to signaling pathways including these kinases[3] demonstrated the role of pathways in the uncontrolled cell growth that is typical of cancer.[4] Later it was realized that many forms of specific signaling proteins (i.e., different related kinases encoded by different genetic loci) existed, and that each member of a family could substitute for another in specific cell types or be aberrantly expressed in some cancers.[5]

In addition, multiple signaling pathways that play important roles in programmed cell death (PI3K-AKT), proliferation (RAS-RAF), cell cycle (Rb-CDK), DNA damage response (P53), and cell adhesion (FAK) were discovered to play roles in cancer etiology.[6] Each pathway, as with the RAS-RAF-MAPK-ERK pathway, contains multiple signaling proteins, with many proteins having known multiple loci encoding related family members. Overall, this creates a situation in which a single aberrant protein (e.g., an oncogene) in a pathway can activate that pathway inappropriately, leading to escape of a cell from a checkpoint on growth. Effectively, each viable cancer therefore has multiple hits (as first proposed by Knudson[7] for the related case of a dominant tumor suppressor), but the hits may be different (i.e., different pathway members) in each cancer, even for cancers of the same apparent phenotype.

Validation of this new view of cancer came with studies of coordinated methylation, mu-

tation, copy number, and expression changes in glioblastoma multiforme and pancreatic cancer.[8–10] In these studies it was demonstrated conclusively that almost all cancers had changes to one protein in each important pathway, but that these proteins were not the same between different individuals. This result suggested that analysis of pathways would be more informative than analysis of genes across a population.

3. High-Throughput Data

Traditional molecular biology and biochemistry involved detailed study of one or a few genes in tightly controlled experimental systems. This approach changed dramatically with the emergence of gene expression microarrays in the mid 1990's.[11,12] These technologies soon allowed researchers to measure levels of mRNA genome-wide and represented the first of many genome-wide measurement technologies. Subsequent advances since the development of microarrays for gene expression have been very rapid. Tiling arrays and array comparative genomic hybridization (aCGH) have allowed increasingly fine-grained measurements of DNA variations. Use of these arrays and custom arrays coupled with immunoprecipitation permit genome-wide measurement of transcription factor and regulatory factor binding. SNPs are now measured genome-wide as well, and SNP-chips also permit estimation of copy number variation (CNV) at increasingly fine resolution. Recently miRNA chips have been developed, so that the abundance of miRNA families can now routinely be measured for all known miRNAs. Coupling microarray technology to methylation-specific precipitation allows measurement of methylation levels in the genome as well. Next-generation sequencing is replacing some of these technologies, now routinely providing genomic-, epigenomic-, and transcript-level measurements. Emerging technologies in nuclear magnetic resonance and mass spectrometry are beginning to provide large-scale measurements of metabolites and proteins, and antibody and reverse-phase protein arrays have the potential to allow genome-wide measurements of protein levels in a microarray format.

As multiple high-throughput measurements representing different molecular entities (e.g., DNA, mRNA, protein) are now routinely made, methods to integrate the data between these different molecular domains are needed. These can be gene-centric, aligning measurements to the genome for instance, or protein-centric, focusing on protein isoforms and including alternative splicing and post-translational modifications.

4. Analysis Approaches and Tools

The simplest approach to account for the heterogeneity introduced by a pathway effect into analysis of high-throughput data is to realize that only a subset of disease samples may harbor a mutation or change in expression and to generate a statistic to address this. In fact, methods to identify these outlier genes have been developed.[13,14] The next step is to generate a pathway or set statistic to replace single gene statistics, which was the focus of methods now known as gene set analysis.[15] However, a model-based analysis that directly utilizes pathway structures to interpret high-throughput data should provide greater power for biological discovery. The modeling methods discussed in the workshop utilized high-throughput measurements of cell lines, model organisms, and tumors to discover novel insights into biological systems.

Cell lines developed from primary tumors have been among the most important tools for discovering the molecular changes underlying cancer and for drug compound screening. A recent study of 30 breast cancer cell lines used expression and proteomics profiles, along with mutational and copy number variation data to build a discrete, rule-based network signaling model for each cell line,[16] based on the Pathway Logic system.[17] Each model has an initial state that represents all expressed proteins in the cell line. Signaling is represented by rule sets, based on experimentally derived protein-protein interactions, which determine a sequence of model states. This approach involves many simplifying assumptions, in particular discretization of data, i.e. each protein component is either present or absent in each state. However, the simplicity makes the model interpretable and it recaptitulates known breast cancer biology and yields useful new hypotheses about aberrant signaling in breast cancer. For example, model analysis elucidates the role of the gene CAV1 in highly aggressive basal B breast cancers and the relationship of PAK1 to MAPK cascade regulation. In particular, the hypothesized importance of PAK1 led to the discovery that PAK1 over-expression may provide a potential clinical marker for the utility of MEK inhibitors in breast cancer treatment.

Genome-scale studies of primary tumors, in increasingly larger patient cohorts, have become widespread. These studies measure multiple biomolecules in tumor tissue and matched normal samples, including gene expression, copy number variation, somatic mutations, SNPs, and methylation level. The volume and complexity of this data requires new analysis methods to reach translational goals, such as improved prognostics and patient-specific therapies. PARADIGM, a probabilistic graphical model that maps multiple patient-specific genome-scale measurements onto curated cancer-related pathways, can be used to infer which components of a pathway (broadly defined as physical entities, gene families, and abstract processes) are activated with respect to a normal cell.[18] This process yields a matrix of integrated pathway activities (IPAs) for each patient. Based on IPA clustering, clinically relevant subgroups of patients were identified, with the potential for improved stratification of patients for targeted therapeutic regimens.

ResponseNet treats genetic library screening results and transcriptional changes measured by microarray experiments within the context of the relationship between signaling protein interactions and transcriptional regulation, integrating multiple types of data (e.g., microarray, genetic library, ChIP-chip) from different experimental sources. It was used to successfully identify pathways involved with α-synuclein toxicity and genes differentially regulated by these pathways.[19] This approach, however, relies on downstream transcriptional changes to drive discovery, and thus can miss important protein interactions changes that do not drive transcriptional change. An alternative approach, an award gathering Steiner tree, was used to identify changes driven by protein interactions in the yeast pheromone response.[20] The Steiner tree was successful in balancing the introduction of false positive interactions from experimental data with the loss of key interactions.

5. Conclusion

Our understanding of biological processes and their control has led to a model of biology in which biological regulatory and metabolic pathways play the dominant role. Evolution has led

to multiple genes in many key families in these pathways, complicating the identification of cell-specific drivers of biological processes. When these drivers are mutated, over-expressed, lost, or replaced by aberrant family members, disease may emerge. Understanding these pathways and identifying the specific members causing disease is critical to elucidating the heterogeneous molecular changes driving disease, identifying subgroups of patients with shared molecular changes, and developing individualized therapies.

References

1. A. Sreekumar, L. M. Poisson, T. M. Rajendiran, A. P. Khan, Q. Cao, J. Yu, B. Laxman, R. Mehra, R. J. Lonigro, Y. Li, M. K. Nyati, A. Ahsan, S. Kalyana-Sundaram, B. Han, X. Cao, J. Byun, G. S. Omenn, D. Ghosh, S. Pennathur, D. C. Alexander, A. Berger, J. R. Shuster, J. T. Wei, S. Varambally, C. Beecher and A. M. Chinnaiyan, *Nature* **457**, 910 (2009).
2. M. H. Cobb, D. J. Robbins and T. G. Boulton, *Curr Opin Cell Biol* **3**, 1025 (1991).
3. G. L. Johnson and R. R. Vaillancourt, *Curr Opin Cell Biol* **6**, 230 (1994).
4. R. Khosravi-Far and C. J. Der, *Cancer Metastasis Rev* **13**, 67 (1994).
5. A. D. Cox and C. J. Der, *Cancer Biol Ther* **1**, 599 (2002).
6. D. Hanahan and R. A. Weinberg, *Cell* **100**, 57 (2000).
7. A. G. Knudson, *Proc Natl Acad Sci U S A* **68**, 820 (1971).
8. S. Jones, X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S. M. Hong, B. Fu, M. T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu and K. W. Kinzler, *Science* **321**, 1801 (2008).
9. D. W. Parsons, S. Jones, X. Zhang, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I. M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, J. Diaz, L. A., J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. Marie, S. M. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu and K. W. Kinzler, *Science* **321**, 1807 (2008).
10. TCGA, *Nature* **455**, 1061 (2008).
11. M. Schena, D. Shalon, R. W. Davis and P. O. Brown, *Science* **270**, 467 (1995).
12. D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton and E. L. Brown, *Nat Biotechnol* **14**, 1675 (1996).
13. J. W. MacDonald and D. Ghosh, *Bioinformatics* **22**, 2950 (2006).
14. R. Tibshirani and T. Hastie, *Biostatistics* **8**, 2 (2007).
15. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church, *Nat Genet* **22**, 281 (1999).
16. L. M. Heiser, N. J. Wang, C. L. Talcott, K. R. Laderoute, M. Knapp, Y. Guan, Z. Hu, S. Ziyad, B. L. Weber, S. Laquerre, J. R. Jackson, R. F. Wooster, W. L. Kuo, J. W. Gray and P. T. Spellman, *Genome Biol* **10**, p. R31 (2009).
17. S. Eker, M. Knapp, K. Laderoute, P. Lincoln, J. Meseguer and K. Sonmez, *Pac Symp Biocomput* , 400 (2002).
18. C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler and J. M. Stuart, *Bioinformatics* **26**, i237 (2010).
19. E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist and E. Fraenkel, *Nat Genet* **41**, 316 (2009).
20. S. S. Huang and E. Fraenkel, *Sci Signal* **2**, p. ra40 (2009).