Omics-Based Discovery Strategies: Collection, Mining and Analysis

David G. Covell, Ph.D.

Screening Technologies Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, NCI-Frederick, Frederick MD 21702, covell@mail.ncifcrf.gov

National Science Foundation's Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery of Innovation

Consistent with the theme of this National Science Foundation's Symposium, a shared challenge to biomedical scientists is the task of organizing the enormous volume of information flowing from omics-based studies to probe disease mechanisms, to identify novel therapeutic targets and to propose markers for designing and monitoring therapeutic studies[1; 2; 3; 4]. The seminal unmet objective underlying this task is that of extracting information from this data sufficient to drive the rational design of therapeutic agents that target specific disease pathways. This conceptual theme offered early motivation for discovering magic bullets, which has now given way to the realization that omics data in general provides only a glimpse of the complexities inherent in these systems [5] and that this data might be, at best, only a weak surrogate for monitoring the underlying biological process. Furthermore, little support exists for a simplified hypothesis developed from the idea that a single entity (gene, nucleic acid or protein) defines a biological outcome; greater support exists for the occurrence of many unanticipated players having desirable and undesirable effects. Despite this dim view, noteworthy successes in linking omics data to drug activity and mechanism of action have offered hope that data mining pursuits may offer more information than initially believed, if collected, mined and analyzed systematically.

Omics-based investigations offer potentially powerful readouts that may be useful for probing the underlying biology of normal and diseased states, identifying novel therapeutic targets and proposing relevant markers for designing therapeutic strategies. A vital component of these investigations involves a systematic analysis of omics data in the context of disease states and small molecules that probe the function of unknown targets responsible for a disease. Numerous systematic data collection strategies are currently underway aimed at the identification of novel, small-molecule, potentially therapeutic, agents that affect a particular disease pathway. Amongst the newcomers is the recent entry of chemical genetics and its efforts to amalgamate disciplines from classical medicinal chemistry and genomics with the wide-ranging new technologies associated with each of these disciplines. Notable in these associated technologies are high throughput screening [6] and methods of target identification and validation [7; 8; 9]. Chemical genetics research is directed at deriving a (biochemical, physiological, pharmacological) understanding of the molecular basis of phenotypes that characterize normal and disease states. Small molecule agents assume the role of molecular probes that can selectively modulate the myriad of interactions that affect and control phenotypes[10]. Readouts from these probes become the basis for interpretations and hypothesis generations about what, if any, components of this vast interacting network may be interpreted in a cause and effect paradigm[11]. Typical approaches appear in high throughput screening efforts that use a library of small molecule against a selected target of interest to observe the effects of target inhibition on phenotypic readout. A hallmark of these efforts is *a priori* knowledge about the target and its potential role in a disease state. These efforts seek 'hits' involving cell- or tissue-based screening, followed by target validation and appropriate specificity checks, leading finally to a network/pathway analysis to understand the basic biology of the molecular target as it functions in a diseased condition. These indirect approaches provide valuable information not produced by classical genetics approaches. Ultimately, the collective body of evidence derived from many cross-disciplines, including chemical genetics, will play contributing roles when validating a therapeutic target.

The NCI Tumor Panel Screening program (referred to hereafter as the NCI60) began in 1990 as a tool for generating a unique readout for small-molecule and natural product extract screening against a selected set of immortalized human cancer cell lines[12; 13; 14].
A variety of analytical approaches using the NCI60 dataset have constructed means to relate similarities in NCI60 bioactivity profiles[15; 16; 17; 18] to reveal the considerable information yet available for mining this unique dataset[16; 17; 19; 20; 21; 22; 23; 24]. All of these efforts support the role of systematic omics analysis to inform the underlying biology of several disease states. Characterizing the generality of this result remains an open challenge, requiring careful examination of each experimental condition. More relevant is the importance of cataloging the genetic snapshot represented by each experimental condition. This information represents the descriptor set necessary for *a priori* assessments that may be useful for connecting a genetic state to the consequences of a chemical perturbation.

Studies utilizing chemicals to perturb biology represent relatively recent efforts intended to contribute towards the urgently needed discovery pipeline [25]. Any effort that efficiently guides a compound into the discovery funnel, by maximizing understanding of and prediction of a compound's mechanism of action, while requiring physicochemical properties necessary for effective and safe compound delivery to the intended target, represents the modern standard of practice. The question still remains as to whether omics approaches are essential for the discovery process, or simply the application of a new technology to an existing problem. Genetic information does appear to move us closer to understanding the complicated pathways and associated pathway products that drive the underlying biology. This degree of complexity has contributed to the movement away from reductionism as a design principle for scientific experimentation (one variable per experiment) towards non-hypothesis driven approaches that effectively create a data pool that requires independent as well as integrative analyses for interpretation

An apparent research bottleneck in this effort is that the tools to analyze complicated datasets require applications of the newest advances in statistics and mathematics [26] [27] and the attraction of young scientists aimed at this effort. A noteworthy complement to analytical strategies is integrated data sharing to virtually extend the number of observations available for analysis [28; 29]. Through global, novel participation it will be possible to gauge the utility of omics-based approaches for probing disease mechanisms, identifying novel therapeutic targets and proposing relevant markers for designing and monitoring therapeutic strategies.

References

1. Berger, A. B., Vitorino, P. M. & Bogyo, M. (2004). Activity-based protein profiling: applications to biomarker discovery, in vivo imaging and drug discovery. *Am J Pharmacogenomics* **4**, 371-81.

2. Phan, J. H., Quo, C. F. & Wang, M. D. (2006). Functional genomics and proteomics in the clinical neurosciences: data mining and bioinformatics. *Prog Brain Res* **158**, 83-108.
3. Kiechle, F. L., Zhang, X. & Holland-Staley, C. A. (2004). The -omics era and its impact. *Arch Pathol Lab Med* **128**, 1337-45.
4. Jay, C., Nemunaitis, J., Chen, P., Fulgham, P. & Tong, A. W. (2007). miRNA profiling for diagnosis and prognosis of human cancer. *DNA Cell Biol* **26**, 293-300.
5. Chung, C. H., Levy, S., Chaurand, P. & Carbone, D. P. (2007). Genomics and proteomics: emerging technologies in clinical cancer research. *Crit Rev Oncol Hematol* **61**, 1-25.
6. Kawasumi, M. & Nghiem, P. (2007). Chemical genetics: elucidating biological systems with small-molecule compounds. *J Invest Dermatol* **127**, 1577-84.
7. Darnell, R. B. (2006). Developing global insight into RNA regulation. *Cold Spring Harb Symp Quant Biol* **71**, 321-7.
8. Wang, Y., Chiu, J. F. & He, Q. Y. (2006). Proteomics approach to illustrate drug action mechanisms. *Curr Drug Discov Technol* **3**, 199-209.
9. Búrdine, L. & Kodadek, T. (2004). Target identification in chemical genetics: the (often) missing link. *Chem Biol* **11**, 593-7.
10. Gangadhar, N. M. & Stockwell, B. R. (2007). Chemical genetic approaches to probing cell death. *Curr Opin Chem Biol* **11**, 83-7.
11. Thorpe, D. S. (2003). Forward & reverse chemical genetics using SPOS-based combinatorial chemistry. *Comb Chem High Throughput Screen* **6**, 623-47.
12. Boyd, M. R. & Paull, K. D. (1995). Some practical considerations and applications of the National Cancer Institute *in vitro* anticancer drug discovery screen. *Drug Development Research* **34**, 91-109.
13. Shoemaker, R. H., Monks, A., Alley, M. C., Scudiero, D. A., Fine, D. L., McLemore, T. L., Abbott, B. J., Paull, K. D., Mayo, J. G. & Boyd, M. R. (1988). Development of human tumor cell line panels for use in disease-oriented drug screening. *Prog Clin Biol Res* **276**, 265-86.
14. Shoemaker, R. H., Scudiero, D. A., Melillo, G., Currens, M. J., Monks, A. P., Rabow, A. A., Covell, D. G. & Sausville, E. A. (2002). Application of high-throughput, molecular-targeted screening to anticancer drug discovery. *Curr Top Med Chem* **2**, 229-46.
15. Rabow, A. A., Shoemaker, R. H., Sausville, E. A. & Covell, D. G. (2002). Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. *J Med Chem* **45**, 818-40.
16. Huang, R., Wallqvist, A., Thanki, N. & Covell, D. G. (2005). Linking pathway gene expressions to the growth inhibition response from the National Cancer Institute's anticancer screen and drug mechanism of action. *Pharmacogenomics J* **5**, 381-99.
17. Huang, R., Wallqvist, A. & Covell, D. G. (2005). Anticancer metal compounds in NCI's tumor-screening database: putative mode of action. *Biochem Pharmacol* **69**, 1009-39.
18. Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer* **6**, 813-23.
19. Bussey, K. J., Chin, K., Lababidi, S., Reimers, M., Reinhold, W. C., Kuo, W. L., Gwadry, F., Ajay, Kouros-Mehr, H., Fridlyand, J., Jain, A., Collins, C., Nishizuka, S.,

Tonon, G., Roschke, A., Gehlhaus, K., Kirsch, I., Scudiero, D. A., Gray, J. W. & Weinstein, J. N. (2006). Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol Cancer Ther* **5**, 853-67.

20. Covell, D. G., Wallqvist, A., Huang, R., Thanki, N., Rabow, A. A. & Lu, X. J. (2005). Linking tumor cell cytotoxicity to mechanism of drug action: an integrated analysis of gene expression, small-molecule screening and structural databases. *Proteins* **59**, 403-33.

21. Huang, R., Wallqvist, A. & Covell, D. G. (2006). Assessment of in vitro and in vivo activities in the National Cancer Institute's anticancer screen with respect to chemical structure, target specificity, and mechanism of action. *J Med Chem* **49**, 1964-79.

22. Huang, R., Wallqvist, A. & Covell, D. G. (2006). Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen. *Genomics* **87**, 315-28.

23. Wallqvist, A., Huang, R., Covell, D. G., Roschke, A. V., Gelhaus, K. S. & Kirsch, I. R. (2005). Drugs aimed at targeting characteristic karyotypic phenotypes of cancer cells. *Mol Cancer Ther* **4**, 1559-68.

24. Wallqvist, A., Huang, R., Thanki, N. & Covell, D. G. (2006). Evaluating chemical structure similarity as an indicator of cellular growth inhibition. *J Chem Inf Model* **46**, 430-7.

25. Muller, G. (2003). Medicinal chemistry of target family-directed masterkeys. *Drug Discov Today* **8**, 681-91.

26. Petsko, G. A. (2006). Do the math. *Genome Biol* 7, 119.

27. Fay, N. (2006). The role of the informatics framework in early lead discovery. *Drug Discov Today* **11**, 1075-84.

28. Jenssen, T. K. & Hovig, E. (2002). The semantic web and biology. *Drug Discov Today* **7**, 992.

29. Neumann, E. K. & Quan, D. (2006). BioDash: a Semantic Web dashboard for drug development. *Pac Symp Biocomput*, 176-87.