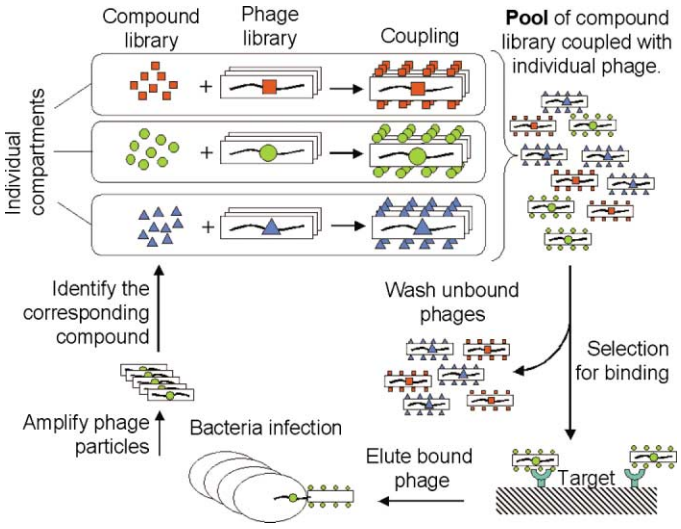

Figure 1. Diagram Illustrating the Formation, Screening, and Analysis of a Small Molecule Library Displayed on Phage

synthesis of the folic acid derivatives. The data analysis is straightforward and easy to follow. In summary, this report presents a compelling case for the utility of the methodology, especially for selecting molecules engaged in cell targeting, uptake, and translocation.

Frederic Fellouse and Kurt Deshayes
Department of Protein Engineering
Genentech Inc.
1 DNA Way
South San Francisco, California 94080

Selected Reading

1. Lenz, G.R., Nash, H.M., and Jindal, S. (2000). Drug Discov. Today *5*, 145–156.
2. Lam, K.S., Lebl, M., and Krchnak, V. (1997). Chem. Rev. *97*, 411–448.
3. Sidhu, S.S., Fairbrother, W.J., and Deshayes, K. (2003). Chembiochem *4*, 14–25.
4. Sandman, K.E., Benner, J.S., and Noren, C.J. (2000). J. Am. Chem. Soc. *122*, 960–961.
5. Woiwode, T.F., Haggerty, J.E., Katz, R., Gallop, M.A., Barrett, R.W., Dower, W.J., and Cwirla, S.E. (2003). Chem. Biol. *10*, this issue, 847–858.
6. Trepel, M., Arap, W., and Pasqualini, R. (2002). R. Curr. Opin. Chem. Biol. *6*, 399–404.
7. Westerhof, G.R., Schornagel, J.H., Kathmann, I., Jackman, A.L., Rosowsky, A., Forsch, R.A., Hynes, J.B., Boyle, F.T., Peters, G.J., and Pinedo, H.M. (1995). Mol. Pharmacol. *48*, 459–471.

Chemistry & Biology, Vol. 10, September, 2003, DOI 10.1016/j.chembiol.2003.09.006

Ontology Recapitulates Physiology

High-content information experiments in the post-genomic era hold the promise of deciphering age-old questions in biology and new ones in the biomedical arena. In response, researchers are devising computationally intensive and novel strategies to extract answers from multidimensional data sets.

The post-genomic era has brought with it a vast collection of data from disparate sources, raising new questions about how to interpret the information and derive something meaningful. First, the human genome sequence and high-density gene expression arrays came, followed by high-throughput bioassays, SNPs, proteome biochips, and, more recently, genome-wide gene knockdown screens in cells, the collective interpretation of which, in the absence of "A Beautiful Mind," is computationally challenging [1–6]. Needless to say, the influx of large-scale data sets has shifted the biomedical research focus toward challenges in computational science [7]. Extracting knowledge contained in the patterns of these experiments into a structured format useful to biologists and medical researchers may highlight an underlying "method to the madness" and could prove critical to an understanding of how cells work.

Attempts to systematically identify novel gene and/or drug function from genome-scale data have thus far relied on acts of heroism both at the bench and in front

of the computer. For example, Hughes et al. collected and assembled a reference database of gene expression profiles from over 300 gene mutations and chemical treatments in yeast and, using a pattern-matching algorithm, were able to assign function to eight novel open reading frames and identify the biochemical target of the topical anesthetic small molecule dyclonine [8]. The key insight here was to generate a reference database with enough samples to extract statistical meaning from the subtle differences between expression profiles. The other insight was to run the experiment in yeast, a genetically tractable organism for which a knockout of each of its ~6000 genes exists. Presumably, to perform the same feat in human cells would require an order of magnitude larger data set and be restricted to a single cell type.

An equally Herculean effort was undertaken by the National Cancer Institute (NCI) in order to bin, by mechanism, cytotoxic small molecules as a function of tumor cell selectivity. Over time, the results have led to an extensive screening database in which measures of growth inhibition ($\log(GI_{50})$) of over 100,000 compounds tested against various subsets of 60–100 tumor cell lines were cataloged. In order to extract meaning from the data set, a relatively new computational tool based on self-organizing maps (SOMs) was implemented to derive testable hypotheses [9]. The mapping strategy allowed compound selectivity patterns to be segregated into highly similar response sets. Then, by analogy to both the patterns and map location of very well characterized, known compounds, novel compounds could be assigned a putative mechanism (i.e., purine biosynthesis, antifolates, apoptosis) and sometimes a precise target family (i.e., topisomerase, cyclin-dependent kinases [CDKs]) [10]. Once again, the key insight that allowed such fine-tuned classification of compound mechanism was the use of a substantial reference database. Both of these studies attest to the idea that a novel, undescribed gene or drug's mechanism of action can be inferred by analogy to a compendium of established data. Because of the significant time and capital expenditure required to create such reference databases, researchers have begun looking for alternatives.

The clear advantage of employing well-described information has recently inspired methods to extract information from undiscovered public knowledge repositories and bibliographic databases, (a.k.a. "free" bases). Sources such as the Medline citation database (http://www.ncbi.nlm.nih.gov/PubMed) of the National Library of Medicine (NLM) and other biomedical indices represent an excellent source for extracting high-density "data" on gene and drug function. Unfortunately, because of wide variations in terminology inherent in archives like Medline compiled over many years from an equally wide range of sources, establishing controlled vocabularies is essential. Stanford University investigators, and recently others, formed the Gene Ontology (GO) Consortium to undertake this onerous "normalization" process and created a critical guide to accurately associate genes with processes, cellular components, and molecular functions [11]. Shortly thereafter, an effort to mine Medline for gene function was exacted by Jenssen et al. in the assembly of PubGene, a full-scale literature network for ~14,000 human genes extracted from titles and abstracts of 10 million Medline records [12]. PubGene is based on the assumption that if two genes are comentioned in a report, there is an underlying biological relationship. Remarkably, despite the obvious caveats (for example, "Gene X does NOT bind to Gene Y" would still register as a positive association), a remarkable enrichment for associated genes was found when compared to the Database of Interacting Proteins (DIP). The main application of PubGene has been to link gene expression profiles to biomedical literature to create "literature gene networks" which, by linking to the MeSH index terms (medical subject headings) such as blood coagulation, inflammation, and chemotaxis, allow assignment of associated gene networks to biological processes.

Analogous efforts in the small molecule realm are being undertaken. In this issue of *Chemistry & Biology*, Root et al. describe a set of chemical and computational tools designed to identify previously unknown associations between mechanism and cellular phenotype [13]. In essence, the authors establish a "mechanism of action" ontology for small molecule compounds: a formal specification to represent compounds and the functional landscape they populate. First, by assembling a collection of well-defined, biologically active compounds (termed ACL, annotated compound library), then assigning to each compound multiple mechanistic, functional descriptors, reinforcing the relevance of these classifications à la Jenssen et al., and calculating the coincidence between compound and descriptor in over 11,000,000 Medline records (termed global mechanism extraction), a substantial reference database for known drugs was realized. The value of this is illustrated experimentally when the authors screen the compound library for antiproliferative activity in A549 human lung carcinoma cells and discover a series of active hits. Predictably, many of the active compounds are associated with tumor or cell death-related terms in Medline; however, a surprising but statistically significant enrichment for the descriptor "ionophore," a term previously unassociated with cell death was uncovered and later verified empirically. By this approach, a previously untested hypothesis, that an ionophore-dependent mechanism might selectively halt A549 tumor cell proliferation, could be made. Thus, the authors have demonstrated the utility of their ontology, when coupled with a selection scheme, for finding novel associations and identifying unanticipated mechanisms contributing to a cellular phenotype. In doing so, the authors have laid the foundation for novel, nondeterministic small molecule mechanism prediction. Taken to its extreme, the expansion of the annotated library and the incorporation of additional descriptors to the ontology will allow more demanding determinations, such as the precise molecular target shared between compounds.

One might ask oneself how much useful information these approaches might afford, which is exactly what informaticians developing analogous methods to extract biomolecular interaction networks, gene regulatory networks, and metabolic pathways from the literature (MeKE, KEGG) are going to find out [14, 15]. In fact, the company Ingenuity was launched to develop knowl-

edge-based systems using natural language-processing techniques and full-text literature mining tools for more comprehensive pathway analysis (Ingenuity, CA). Clearly, advances in text mining of factual and literature databases are extending the knowledge base further, demonstrating how text-based information repositories can be used as multidimensional data troves for deciphering genomic scale experiments.

Jeremy S. Caldwell
Department of Drug Discovery
Genomics Institute of the Novartis Research Foundation
San Diego, California 92121

Selected Reading

1. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). Science *291*, 1304–1351.
2. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhardt, D.J. (1999). Nat. Genet. *21*, 20–24.
3. Sundberg, S.A. (2000). Curr. Opin. Biotechnol. *11*, 47–53.
4. Jurinke, C., van den Boom, D., Cantor, C.R., and Koster, H. (2002). Adv. Biochem. Eng. Biotechnol. *77*, 57–74.
5. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. (2001). Science *293*, 2101–2105.
6. Ziauddin, J., and Sabatini, D.M. (2001). Nature *411*, 107–110.
7. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999). Proc. Natl. Acad. Sci. USA *96*, 2907–2912.
8. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. (2000). Cell *102*, 109–126.
9. Sherlock, G. (2000). Curr. Opin. Immunol. *2*, 201–205.
10. Rabow, A.A., Shoemaker, R.H., Sausville, E.A., and Covell, D.G. (2002). J. Med. Chem. *45*, 818–840.
11. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Nat. Genet. *25*, 25–29.
12. Jenssen, T.K., Laegreid, A., Komorowski, J., and Hovig, E. (2001). Nat. Genet. *28*, 21–28.
13. Root, D.E., Flaherty, S.P., Kelley, B.P., and Stockwell, B.R. (2003). Chem. Biol. *10*, this issue, 881–892.
14. Chiang, J.H., and Yu, H.C. (2003). Bioinformatics *19*, 1417–1422.
15. Ogata, H., Goto, S., Fujibuchi, W., and Kanehisa, M. (1998). Biosystems *47*, 119–128.