DOI: 10.1002/anie.200604972

Small-Molecule Profiles

Understanding the Activity of Small Molecules in the Genomics Context**

*Jun-Seok Lee and Young-Tae Chang**

Keywords:
analytical methods · connectivity map ·
gene expression · genomics · molecular profiling

The chemical genetics approach that utilizes small molecules to modulate biological pathways plays a central role in the field of drug discovery and development. However, owing to the complexity of biological systems, determination of the biological activity of each molecule and identification of targets require a number of tedious experiments. This issue has motivated the development of methods for high-throughput analysis of the activity of small molecules. Looking back over a decade of research, our understanding about the function of small molecules with relation to disease and gene expression has improved vastly. Recently, a new promising "connectivity map" approach in the small-molecule profiling method was reported by Golub and co-workers.[1–3] Among the approaches tried so far, the connectivity map demonstrates the most powerful potential to aid our understanding of drug–drug, drug–gene, and drug–disease relationships.

Before the development of the connectivity map, the US National Cancer Institute (NCI) 60 drug screen (NCI60) was the most extensive platform for systematic studies of the activity of small molecules. Since its development in the late 1980s, a diverse range of synthetic small molecules and natural products were screened in a variety of cancer cell lines, and those data were used to extract valid information for the mode of action of the tested compound.[4,5] In the NCI approach, patterns of activity for small molecules were generated by measuring cytotoxicity levels in 60 different cancer cell lines and were analyzed by heat maps (correlation coefficient maps) and hierarchical clustering.[6] However, analysis of molecular profiles using GI50 patterns in NCI60 has an implicit limitation as the pattern is based on only one final phenotypic response in the cellular system. Overall patterns may have coordinated fingerprints for each compound, but these are not directly connected to the biological pathway.

In 2000, Hughes et al. reported the compendium approach, which uses gene-expression signatures for profiling instead of cytotoxicity in cells.[7] The authors suggested that within the whole gene-expression profile lies a cellular state's compendium, and they proved the concept of this approach at least in yeast. Encouraged by these former studies, Golub and co-workers developed a more robust system, the so-called connectivity map, which involves analysis of the gene-expression signature of small molecules in mammalian cells by utilizing a novel data analysis method.[1]

The technical part of this connectivity map system is simple. The authors collected 164 distinct small molecules. Each of these molecules was incubated for 6 h at a suitable concentration (mostly 10 μM) in mammalian cells (mainly MCF7 breast cancer cells). Compared with each vehicle and treatment gene-expression pattern, the activity of each compound was represented by the rank-ordered list of 22283 genes. Once the reference database collection was completed (453 individual examples), the authors input a query to calculate the connectivity score. The query signature and the relationships of the activities of the reference small molecules were presented by a bar view image, which is a rank-ordered list of reference compounds sorted by connectivity score (Figure 1).

The connectivity map has five apparent merits:

1) It uses the gene-expression signature as a profiling probe. This has several benefits compared to the NCI method. The mRNA-expression level is directly related to the biological pathway, and the gene-expression profile itself can be used for functional gene studies. Despite these benefits, gene-expression profiling data are not easy to handle owing to the difficulties of massive data analysis.[8]
2) The pattern analysis involves non-parametric, Kolmogorov–Smirnov statistical rank-based pattern matching. The major difference between the conventional hierarchical method and the new method (gene set enrichment analysis, GSEA[9]) is that the former uses the fold change values in the analysis, while the latter utilizes the rank in the gene list. Most conventional methods for microarray data pattern analysis focus on genes that display dramatic fold changes, and the fold change values are directly used for pattern recognition.[10] The fold change values may vary in each experiment, so the clustering results suffer from experimental deviation. By using a rank-

[*] J.-S. Lee, Prof. Dr. Y.-T. Chang
Department of Chemistry
New York University
New York, NY 10003 (USA)
Fax: (+1) 212-995-4203
E-mail: yt.chang@nyu.edu

[**] J.-S.L. was supported by a Korea Research Foundation Grant funded by the Korean Government MOEHRD Basic Research Promotion Fund (KRF-2005-C00088).

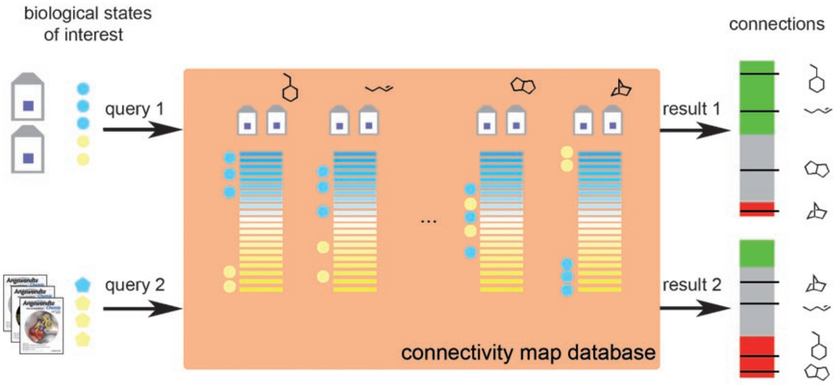

Figure 1. Representation of the connectivity map query process (modified slightly from the original figure[1]). A query signature can be generated not only from microarray experiments (query 1) but also from published gene-regulation pattern information (query 2).

ing system instead of the absolute fold change, the noise induced by differences in experimental conditions is dramatically reduced in the connectivity map, which allows comparison of data sets from different experimental sets. Moreover, there is no standard way of estimating the significant decision line in hierarchical methods. For example, in a hierarchical clustering heat map, it is hard to say where the borderline for positive correlation lies. In contrast, the connectivity map always gives a connectivity score. The sign of the connectivity score represents whether two signatures have positive or negative correlations (null means no correlation).

3) A query to the connectivity map database is independent of the experimental platform. A query signature needs only two gene set lists (up-regulated gene set and down-regulated gene set). The results from not only microarray experiments but also wet biological experiments can make a query.
4) This methodology can form a broad range of connections among small molecules, genes, and diseases in terms of gene-expression profiles. This is a significant conceptual extension from NCI analysis, which allows only a comparison of compounds.
5) Finally, all of these reference data and analysis techniques are accessible through the internet (http://www.broad.mit.edu/cmap/). While this latter point is more about the authors' policy rather than the technology itself, the generality of the connectivity map, which embraces many different experimental formats, and interactive open resources are important merits from a practical point of view.

The authors evaluated the performance of the connectivity map by asking two questions: 1) Can the connectivity map properly predict small-molecule activity? 2) Can this model system find connection between disease and small molecules?

Regarding the first question, their primary data set showed high correlation among three known histone deacetylase (HDAC) inhibitors and also a correlation between estrogen receptor (ER) agonists and antagonists. These results indicate that the gene-expression signature can be an outstanding probe for profiling molecular activity. However, for the broad range of potential users, there are still many other factors to consider. As the authors mentioned in their report, the incubation time, assay cell type, and drug dose concentrations can significantly affect the gene-expression signature. Specifically, several different cell types have different contents of protein expression. Although the signatures of HDAC inhibitors from five different cell lines have a high connectivity score, in some cases, different cell line conditions do affect the molecular activity profile, as shown in the case of the ER ligand signature in PC3 and HL60 cells, which do not express ER.[1] The authors pointed out that the connectivity score for estradiol (an ER agonist) in ER-expression-free cell lines showed the need for an expanded reference data set.

Issues of drug dosage and incubation time are also complicated. Chemically induced changes in intercellular signal pathways are dynamic phenomena. Microarray-based studies on drug dose and incubation time indicated that one small molecule can have different gene-expression signatures depending on incubation time and dose.[11,12] Therefore, more efforts are required to optimize the conditions for a universal signature (Table 1).

To answer the second question, it was noted that one unique feature of the connectivity map is the open query system. The authors tried to collect genetic study data from known diseases and generated query signatures to disease models such as diet-induced obesity and Alzheimer's disease (AD). AD queries from two independent sources gave statistically significant negative connectivity with 4,5-dianilinophthalimide (DAPH) treatment. In an independent study, DAPH was reported as an agent that specifically decreases the β-sheet content of aggregating Aβ1-42 peptides in vitro.[13] Golub and co-workers' result supports the fact that DAPH could be a novel lead compound for drugs for the treatment of AD. Further practical applications of the connectivity map led also to the identification of a novel class of heat-shock-protein 90 (HSP90) pathway modulators (gedunin and celastrol)[2] and the drug-resistance action of rapamycin in acute lymphoblastic leukaemia.[3] These examples show how the connectivity map can facilitate the prediction of targets in chemical genetics by using the gene-expression signature.

The contemporary versions of the connectivity map's query signature do not request any special "form", such as a minimum number of tags, and thus allows full flexibility for diverse purposes. Individual research groups can develop their own application method using the given database. In addition, the connectivity map is not limited by experimental platform and all databases are freely accessible to the public. However, to increase the signature and profiling database of small molecules,

Table 1: Comparison of the NCI60 method and the connectivity map.

Features	NCI60 method	Conventional gene-expression profile	Connectivity map
Expected result	comparison of compounds	comparison of compounds, identification of signaling pathway	identification of drug–gene-disease connection and signaling pathway
Analysis parameter	cytotoxicity	various gene arrays	22 283-gene expression profile (Affymatrix array)
Query source	GI50 data from 60 cancer cells	fold change in gene expression	pair of gene lists (up and down) from various formats
Analysis method	hierarchical analysis	hierarchical analysis	nonparametric, rank-based statistical analysis
Advantage	extensive reference data accumulated	information-rich	low noise
Disadvantage	limited information	batch-to-batch variation (high noise)	more validation required

highly collaborative efforts from the community will be necessary. Challenges that remain include the collection of extensive drug-screening data and the determination of dose/time dependency and cell-type specificity. Once established and popularized, the connectivity map will provide a universal language to better understand the activity of small molecules in the context of genomic.

Published online: April 3, 2007

[1] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, T. R. Golub, *Science* **2006**, *313*, 1929–1935.

[2] H. Hieronymus, J. Lamb, K. N. Ross, X. P. Peng, C. Clement, A. Rodina, M. Nieto, J. Du, K. Stegmaier, S. M. Raj, K. N. Maloney, J. Clardy, W. C. Hahn, G. Chiosis, T. R. Golub, *Cancer Cell* **2006**, *10*, 321–330.

[3] G. Wei, D. Twomey, J. Lamb, K. Schlis, J. Agarwal, R. W. Stam, J. T. Opferman, S. E. Sallan, M. L. den Boer, R. Pieters, T. R. Golub, S. A. Armstrong, *Cancer Cell* **2006**, *10*, 331–342.

[4] K. D. Paull, R. H. Shoemaker, L. Hodes, A. Monks, D. A. Scudiero, L. Rubinstein, J. Plowman, M. R. Boyd, *J. Natl. Cancer Inst.* **1989**, *81*, 1088–1092.

[5] J. N. Weinstein, K. W. Kohn, M. R. Grever, V. N. Viswanadhan, L. V. Rubinstein, A. P. Monks, D. A. Scudiero, L. Welch, A. D. Koutsoukos, A. J. Chiausa, K. D. Paull, *Science* **1992**, *258*, 447–451.

[6] J. N. Weinstein, T. G. Myers, P. M. O'Connor, S. H. Friend, A. J. Fornace, Jr., K. W. Kohn, T. Fojo, S. E. Bates, L. V. Rubinstein, N. L. Anderson, J. K. Buolamwini, W. W. van Osdol, A. P. Monks, D. A. Scudiero, E. A. Sausville, D. W. Zaharevitz, B. Bunow, V. N. Viswanadhan, G. S. Johnson, R. E. Wittes, K. D. Paull, *Science* **1997**, *275*, 343–349.

[7] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, S. H. Friend, *Cell* **2000**, *102*, 109–126.

[8] D. E. Bassett, Jr., M. B. Eisen, M. S. Boguski, *Nat. Genet.* **1999**, *21*, 51–55.

[9] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15 545–15 550.

[10] D. K. Slonim, *Nat. Genet.* **2002**, *32 Suppl*, 502–508.

[11] Y. Zhou, F. G. Gwadry, W. C. Reinhold, L. D. Miller, L. H. Smith, U. Scherf, E. T. Liu, K. W. Kohn, Y. Pommier, J. N. Weinstein, *Cancer Res.* **2002**, *62*, 1688–1695.

[12] J. N. Weinstein, Y. Pommier, *Nat. Biotechnol.* **2006**, *24*, 1365–1366.

[13] B. J. Blanchard, A. Chen, L. M. Rozeboom, K. A. Stafford, P. Weigele, V. M. Ingram, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 14 326–14 332.