Molecular Systems Biology (2005) doi:10.1038/msb4100013
© 2005 EMBO and Nature Publishing Group All rights reserved 1744-4292/05
www.molecularsystemsbiology.com
Article number: 2005.0009

molecular systems biology

NEWS AND VIEWS

Extracting functional and regulatory order from microarrays

Sandrine Imbeaud and Charles Auffray

Array s/IMAGE, Genexpress, Functional Genomics and Systems Biology for Health, LGN—UMR 7091—CNRS and Pierre and Marie Curie University of Paris 6, Villejuif, France

Molecular Systems Biology 25 May 2005; doi:10.1038/msb4100013

Systems approaches for understanding biological complexity and studying diseases rely on iterative and extensive characterization of genes, transcripts, proteins and their interactions, generation of hypotheses about how they functionally inter-relate within subsystems, conversion of these hypotheses into formal mathematical models and their experimental testing (Auffray *et al*, 2003a). In this context, modeling of gene regulatory networks from functional annotations is currently performed top-down, studying global network architecture and performance (Bray, 2003), and bottom-up, identifying modular subsystems from functional genomics data (Alon, 2003).

Because transcription is the first step of gene expression subjected to extensive regulations by internal and external factors, systems approaches rely heavily on gene expression data. Microarray technology has developed steadily for three decades to allow measurements of expression levels for thousands of genes in different biological contexts, and a wealth of such data is now available in public repositories (Ball *et al*, 2004). The expectation is that microarray analysis will help elucidating what the genes do, when, where and how they are expressed as elements of an orchestrated system under the effects of perturbations, and thus reveal the underlying transcriptional regulatory networks.

Two recent papers report attempts to develop an optimized framework for functional annotation and reconstruction of regulatory networks using large-scale expression data sets combined with protein interaction and phenotypic data in yeast (Tanay *et al*, 2005; Zhou *et al*, 2005). These approaches are designed to identify genes with similar functions, but not necessarily coexpressed, and to extract essential features of regulatory networks through analysis of independent data sets. Zhou *et al* first identified a collection of coexpressed gene pairs (doublets) representing functional modules in individual data sets, based on expression correlation and functional annotation, most of which were functionally homogeneous. Then, using this first-order meta-information, they conducted a second-order expression analysis, assembling pairs of doublets (quadruplets) found tightly coregulated across multiple data sets into context-dependent regulatory modules. Similarly, Tanay *et al* used biclustering within a large microarray data compendium to identify relevant functional modules. These approaches generated functional predictions consistent with experimental studies, identified novel cross-doublet gene pairs missed in the standard analyses and allowed assignment of novel functions to a number of previously uncharacterized genes. These achievements represent significant improvements compared to previous studies, since only half of the globally coexpressed gene pairs identified by the standard methods are functionally homogeneous, and analysis of a compendium of yeast expression profiles yielded only a handful of functional assignments (Hughes *et al*, 2000; Auffray *et al*, 2003b).

In addition, Zhou *et al* assembled gene regulatory networks by using first-order expression correlation of target gene modules as an activity profile for the transcriptional factor regulating them, and second-order expression correlations between the activity profiles of transcriptional modules to measure cooperativity between transcription factors. Through integration with protein–protein and protein–DNA interaction data, functionally consistent transcription modules controlled by distinct transcription factors and displaying high second-order correlations were shown to participate in transcription cascades. Thus high-order clustering of transcriptional modules identifies potential interconnectivity between groups of genes participating in the same biological processes, and provides indirect assignment of transcription factors to these processes, overcoming their low expression levels. This represents another improvement over conventional approaches, which are limited by their inability to reconstruct the hidden organization of the regulatory signals (Wei *et al*, 2004): high-order analyses go one step further to capture combinatorial coregulations for genes that do not exhibit identical expression patterns.

Despite the significant progress that such data-driven network assembly methods represent, due to the underlying network complexity, it remains extremely difficult to reconstruct complete regulatory networks exclusively based on the information available from microarrays, even when combined with other types of data in higher order analyses (Wei *et al*, 2004; Papin *et al*, 2005). This is currently limiting our ability to understand the biological significance of the topological properties of the reconstructed networks, which have typically scale-free and small-world architectures (Grigorov, 2005). Combinatorial expansion in the number of potential network structures and comprehensive evaluation of their consistency are key challenges that the approach developed by Zhou *et al* does not entirely overcome, particularly since the number of possible alternate genetic regulatory networks highly depends on the size and type of the data sets and the maximum number of regulatory inputs per gene (Orrell *et al*, 2005). Using a Bayesian modeling approach imposing severe constraints on network architecture, several groups have successfully overcome some of these limitations, inferring transcriptional regulatory modules through a high-order analysis of microarray data combined with genotyping and phenotypic data in recombinant inbred mice (Bystrykh *et al*, 2005; Chesler *et al*, 2005; Hubner *et al*, 2005; Li *et al*, 2005).

However, a great deal of biological information is most likely contained in the absolute expression levels, including the large number of those of low magnitude that are subject to chaotic fluctuations and trigger the emergence of self-organization in complex biological systems (Auffray *et al*, 2003b). Such fluctuations are unlikely to be captured by high-order expression analysis when it only considers functional links that are simultaneously turned on or off over various conditions, and is limited by the current inability of high-throughput technologies to provide the accurate and consistent data required (Jarvinen *et al*, 2004). Due to insufficient standardization in experiment description, including array element description and annotation, and irregularities in data integrity (Brazma *et al*, 2001; Grunenfelder and Winzeler, 2002), microarrays represent an incompletely mature technology using a variety of platforms and analysis tools, which are often difficult to compare. Thus, poorly documented variations exist within any given microarray data set, especially when different generations of microarrays are considered together (Hwang *et al*, 2004; Shi *et al*, 2004). They are likely to influence significantly both first-order and high-order analyses, as shown by the influence of RNA integrity on expression level measurements (Imbeaud *et al*, 2005). Such variations should therefore be documented using vigilant experimental and data processing pipelines rather than masked, as is currently done in most microarray studies.

References

Alon U (2003) Biological networks: the tinkerer as an engineer. *Science* **301:** 1866–1867

Auffray C, Imbeaud S, Roux-Rouquie M, Hood L (2003a) From functional genomics to systems biology: concepts and practices. *CR Biol* **326:** 879–892

Auffray C, Imbeaud S, Roux-Rouquie M, Hood L (2003b) Self-organized living systems: conjunction of a stable organization with chaotic fluctuations in biological space-time. *Philos Transact A* **361:** 1125–1139

Ball C, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M, Sansone SA, Sherlock G, Spellman P, Stoeckert C, Tateno Y, Taylor R, White J, Winegarden N (2004) Standards for microarray data: an open letter. *Environ Health Perspect* **112:** A666–A667

Bray D (2003) Molecular networks: the top-down view. *Science* **301:** 1864–1865

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:** 365–371

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37:** 225–232

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* **37:** 233–242

Grigorov MG (2005) Global properties of biological networks. *Drug Discov Today* **10:** 365–372

Grunenfelder B, Winzeler EA (2002) Treasures and traps in genome-wide data sets: case examples from yeast. *Nat Rev Genet* **3:** 653–661

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Muller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* **37:** 243–253

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000) Functional discovery via a compendium of expression profiles. *Cell* **102:** 109–126

Hwang KB, Kong SW, Greenberg SA, Park PJ (2004) Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics* **5:** 159

Imbeaud S, Graudens E, Boulanger V, Barlet X, Zaborski P, Eveno E, Mueller O, Schroeder A, Auffray C (2005) Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Res* **30:** e56

Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O (2004) Are data from different gene expression microarray platforms comparable? *Genomics* **83:** 1164–1168

Li H, Lu L, Manly KF, Chesler EJ, Bao L, Wang J, Zhou M, Williams RW, Cui Y (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum Mol Genet* **14:** 1119–1125

Orrell D, Ramsey S, de Atauri P, Bolouri H (2005) A method for estimating stochastic noise in large genetic regulatory networks. *Bioinformatics* **21:** 208–217

Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* **6:** 99–111

Shi L, Tong W, Goodsaid F, Frueh FW, Fang H, Han T, Fuscoe JC, Casciano DA (2004) QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Rev Mol Diagn* **4:** 761–777

Tanay A, Steinfeld I, Kupiec M, Shamir R (2005) Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Mol Syst Biol* 29 March 2005; doi:10.1038/msb4100005

Wei GH, Liu DP, Liang CC (2004) Charting gene regulatory networks: strategies, challenges and perspectives. *Biochem J* **381:** 1–12

Zhou XJ, Kao MC, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WH (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol* **23:** 238–243