Molecular Systems Biology 3; Article number 77; doi:10.1038/msb4100118
Citation: *Molecular Systems Biology* 3:77
© 2007 EMBO and Nature Publishing Group All rights reserved 1744-4292/07
www.molecularsystemsbiology.com

NEWS AND VIEWS

Size matters: network inference tackles the genome scale

Boris Hayete[1], Timothy S Gardner[1,2] and James J Collins[1,2]

[1] Bioinformatics Program and Center for BioDynamics, Boston University, Boston, MA, USA and [2] Department of Biomedical Engineering, Boston University, Boston, MA, USA

Molecular Systems Biology 13 February 2007; doi:10.1038/msb4100118

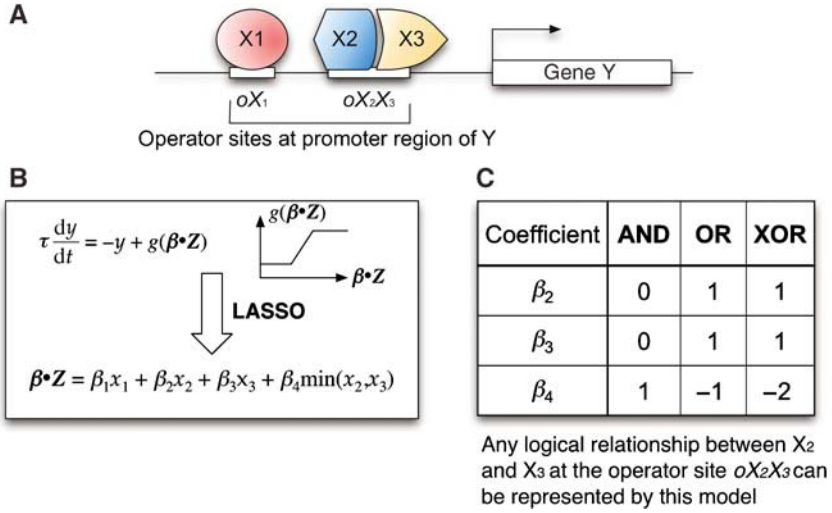
The growing importance of microarray data challenges biologists, and especially the systems biology community, to come up with genome-scale analysis methods that can convert the large quantity of available high-throughput data into high-quality systems-level insights. One area of systems-level analysis that has received considerable attention in recent years is that of inferring molecular-level regulation, with frequent focus on transcriptional regulatory networks (Kholodenko *et al*, 1997; Tavazoie *et al*, 1999; Gardner *et al*, 2003; Segal *et al*, 2003; Beer and Tavazoie, 2004; Yu *et al*, 2004; di Bernardo *et al*, 2005; Gardner and Faith, 2005; Woolf *et al*, 2005; Margolin *et al*, 2006; Faith *et al*, 2007). As microarrays provide a tool for measuring transcript levels of the whole genome, recent interest has shifted to inferring networks on a genome scale. The less-studied organisms are a natural starting point for such mapping, as it is for these organisms that the rapid, genome-scale identification of regulatory structure is most needed.

In a recent study, Bonneau *et al* (2006) apply the Inferelator, their elegant new algorithm, for inferring gene networks, to precisely such a little-studied but important organism. Specifically, the authors focus on *Halobacterium NRC-I*, a model archaeon (DasSarma *et al*, 2006), to show that, at least for a small genome, it is possible to determine a sizeable portion of the transcriptional regulatory network from microarrays without much prior knowledge. This choice of an organism has two practical advantages. First, the salt-loving *NRC-I* is one of a handful of *Halobacteria* for which transformation techniques have been well studied, allowing *in vivo* validation of network predictions. Second, *NRC-I*'s genome is relatively small and thus, its regulation ought to be comparatively easy to reconstruct. Small genome or not, putting high-throughput profiling technologies to work on the genome scale requires a confluence of robust algorithms, biologically plausible simplifying assumptions, and a robust verification strategy. The work of Bonneau *et al* (2006) is a good example, using multiple tools in the bioinformatics toolbox to build a credible blueprint of a transcriptional-regulatory network involving thousands of genes and more than 100 transcription factors.

In order to appreciate the need for a well-structured approach to regulatory mapping, consider the mathematical and biological scope of this cross-disciplinary problem. The tiny archaeon *Halobacterium NRC-I* contains about 2400 genes. For each one of these, the goal is to understand the transcriptional regulatory apparatus—that is about 2400 question marks, each with thousands of possible answers in the form of a set of transcriptional regulators. Put that against a typical compendium size of several hundred chips for a given organism, and you get what is known as a 'small n, large p' problem, where the number of possible parameters (regulators), p, dwarfs the number of data points (microarrays), n, available to define them. This problem gets considerably worse for complex organisms, where a larger number of available microarrays are more than offset by the vast complexity of large genomes, alternate splice variants, and multiple layers of regulation. For network inference algorithms, 'small n, large p' means dearth of data and very high computational demands.

As if this computational complexity were not bad enough, there is the inherent high dimensionality in the biological realm. Regulation happens in the domains of mRNA, proteins, metabolites, kinases, acetylases, and so on, and through a variety of pleiotropic perturbations and influences, such as salinity, temperature, and cell-wall permeability. As the best high-throughput data capture only mRNA, one must make simplifying assumptions and skip many important parameters. Bonneau and colleagues' best simplifying assumption is to focus on predicting the targets of transcription factors in the network, along with some key environmental influences. When only transcription factors are allowed to regulate other genes, the 'p' in the 'small n, large p' problem is no longer so big. In fact, at 120, it is smaller than the number of chips (268) used in this study.

To further constrain the network learning problem, the Inferelator performs a pre-processing step of bi-clustering—organizing experimental data by both genes and conditions. This algorithm, the cMonkey (Reiss *et al*, 2006), allows further reduction of dimensionality by collapsing genes into conditionally coexpressed modules. cMonkey identified 300 such bi-clusters, and 159 individual genes that could not be grouped, a nearly six-fold reduction in dimensionality. Crucially, as the composition of the culture medium used for the microarray-profiled experiments is known, each bi-cluster's grouping of genes by experimental condition suggests plausible metabolic or environmental effectors of regulation. The authors exploit this benefit of their approach in one of their verifying experiments. Bi-clustering, therefore, serves two ends: it limits the number of genes, and thus variables to reconstruct, to fewer than 500 (including only 80


A
X1
X2
X3
Gene Y
oX_1
oX_2X_3
Operator sites at promoter region of Y
B
$g(\beta\bullet Z)$
$\beta\bullet Z$
LASSO
C
Any logical relationship between X_2 and X_3 at the operator site oX_2X_3 can be represented by this model

$$\tau\frac{dy}{dt} = -y + g(\beta\bullet Z)$$

$$\beta\bullet Z = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \min(x_2, x_3)$$

Coefficient	AND	OR	XOR
β_2	0	1	1
β_3	0	1	1
β_4	1	−1	−2

Figure 1 (**A**) Schematic diagram of a hypothetical bacterial operon, represented by a single gene *Y*, which is regulated by a protein X_1 and a protein complex X_2X_3. (**B**) Within its dynamic range, the level of the transcript *y* may be modeled as a function of transcripts of the regulatory proteins X_1, X_2, and X_3. The *min* function captures the notion of cooperativity, and the general form of *g* incorporates saturation effects. On the genome scale, the initial model for regulation of *y* would involve all possible transcription factors, and would greatly benefit from parameter shrinkage by LASSO. (**C**) This table illustrates the representative power of the chosen design matrix. The model can capture AND, OR, and XOR logical functions and saturation effects (not shown). Assigning the shown values to the coefficients from (B) would cause the model to represent the corresponding logical function for the interaction of X_2 and X_3.

TFs and metabolites), and places each predicted regulatory interaction into an experiment-specific context.

The problem now becomes mathematically well-posed, and the authors solve it using LASSO regression, a sparse regression method designed just for such computationally difficult problems (Tibshirani, 1996). LASSO works by selecting a small set of the most likely regulators of a given gene, and simultaneously determines a quantitative influence function relating regulator expression to target expression (Figure 1). In addition, the authors extend the LASSO algorithm beyond its typical linear domain by including piecewise and nonlinear terms in the regression to model saturation effects and pairwise combinatorial regulation. With this approach, the authors construct a model of transcription regulation in *Halobacterium* that matches 80 transcription factors to 500 predicted gene targets and captures the putative metabolic controllers of these pathways. This is an impressive result, both in size and regulatory complexity, particularly in light of the relatively modest size of the experimental data set (i.e., 268 microarrays). Moreover, this represents a dramatic leap in our understanding of this little-studied organism.

Having obtained the first-pass transcriptional blueprint, Bonneau and colleagues ask the obligatory next question: how much do we trust this network? In network inference, three broad types of verification are possible: computational verification through cross-validation, *in vivo* verification, and literature-driven curation. To be effective, the last approach should leverage a large data set documenting connectivity known in the literature, such as TransFac (Matys *et al*, 2003) or RegulonDB (Salgado *et al*, 2006). This type of verification not being available for *Halobacterium*, the authors vigorously pursue the former two, including knockout experimentation and ChIP-chip analysis, demonstrating that their network can serve as a reliable and useful blueprint of *Halobacterium NRC-I*'s transcriptional regulation.

Bonneau *et al* (2006) show the feasibility of mapping a genome-scale regulatory network from a modestly sized compendium of microarrays, an important success for the systems biology community. As microarray technology continues to improve and costs drop, growing databases of microarrays present an opportunity to infer ever more complex regulatory networks in both microbes and higher organisms. Abundance of data fuels the need for a network inference case study that would clearly map the boundaries of what is possible with today's network mapping algorithms. To this end, we believe that the once and future model organisms like *Escherichia coli* and *Saccharomyces cerevisiae*, buoyed by extensive bodies of literature and large databases such as RegulonDB, SGD (Christie *et al*, 2004), and TransFac, may represent attractive short-term targets for network inference studies. In addition to the use of curated data sets, it may be possible to seed organisms with small synthetic *in vivo* networks, the connectivity of which is known by design, and to measure the success of network reconstruction on the whole by success or failure to reconstruct the seed. We are aware of at least one lab doing such work (Cantone *et al*, 2006). Biological yardsticks in general will gain in importance, as they supplement *in silico* testing and usher in algorithms' transition from design to practical use, and from simple organisms to higher eukaryotes.

Challenges remain, but we see the immediate future of network inference as promising and bright. Molecular biologists have long been looking for ways to generate more oomph from their microarrays. Systems biology may have some answers, and we laud Bonneau and colleagues for providing an illuminating step in that direction.

References

Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. *Cell* **117:** 185

Bonneau R, Reiss D, Shannon P, Facciotti M, Hood L, Baliga N, Thorsson V (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol* **7:** R36

Cantone I, di Bernardo D, Cosma MP (2006) Benchmarking reverse-engineering strategies via a synthetic gene network in *Saccharomyces cerevisiae*. *DIMACS Workshop on Dialogue on Reverse Engineering Assessment and Methods (DREAM)*. New York, NY:Wave Hill Conference Center

Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM (2004) *Saccharomyces* genome database (sgd) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* **32** (Suppl 1)**:** D311–D314

DasSarma S, Berquist B, Coker J, DasSarma P, Muller J (2006) Post-genomics of the model haloarchaeon halobacterium sp. Nrc-1. *Saline Systems* **2:** 3

di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotech* **23:** 377

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**: e8; doi:10.1371/journal.pbio.0050008

Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301:** 102–105

Gardner TS, Faith JJ (2005) Reverse-engineering transcription control networks. *Phys Life Rev* **2:** 65

Kholodenko BN, Hoek JB, Westerhoff HV, Brown GC (1997) Quantification of information transfer via cellular signal transduction pathways. *FEBS Lett* **414:** 430

Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform* **7** (Suppl 1)**:** S7

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31:** 378

Reiss D, Baliga N, Bonneau R (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinform* **7:** 280

Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J (2006) Regulondb (version 5.0): *Escherichia coli* k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* **34** (Suppl 1)**:** D394–D397

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34:** 166

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* **22:** 281

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* **58:** 288

Woolf PJ, Prudhomme W, Daheron L, Daley GQ, Lauffenburger DA (2005) Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics* **21:** 741–753

Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED (2004) Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20:** 3594–3603