Review

CellPress

Modeling genomic regulatory networks with big data

Hamid Bolouri

Division of Human Biology, Fred Hutchinson Cancer Research Center (FHCRC), 1100 Fairview Avenue North, PO Box 19024, Seattle, WA 98109, USA

High-throughput sequencing, large-scale data generation projects, and web-based cloud computing are changing how computational biology is performed, who performs it, and what biological insights it can deliver. I review here the latest developments in available data, methods, and software, focusing on the modeling and analysis of the gene regulatory interactions in cells. Three key findings are: (i) although sophisticated computational resources are increasingly available to bench biologists, tailored ongoing education is necessary to avoid the erroneous use of these resources. (ii) Current models of the regulation of gene expression are far too simplistic and need updating. (iii) Integrative computational analysis of large-scale datasets is becoming a fundamental component of molecular biology. I discuss current and near-term opportunities and challenges related to these three points.

Gene regulatory networks (GRNs)

The past few years have witnessed dramatic milestones in high-throughput sequencing, large-scale data generation, cloud computing, and computational biology. Supra-exponential improvements in the throughput and cost of DNA sequencing (http://www.genome.gov/sequencingcosts/) have been accompanied by improvements in accuracy and reductions in the required sample size. These improvements have in turn led to the widespread adoption of a broad range of sequencing-based technologies (reviewed in [1]) to characterize not only genomes but also the regulatory interactions that allow genomes to specify cellular structure, function, and behavior.

GRNs are defined as the set of interactions among genes and their products (RNAs and proteins) that determine the isoforms, location (cell type), timing, and rate of RNA expression [2] (see Figure 1 for examples). With the possible exception of some metabolic and physiological processes, GRNs are the primary drivers of cellular behavior and function.

Because GRNs are ultimately specified by the digital code of DNA, they are uniquely accessible to both high-throughput sequencing-based technologies and to computational modeling and analysis. At the same time, GRNs are both complex (i.e., can exhibit hard-to-predict/nonlinear behaviors) and complicated (i.e., they are composed of large numbers of component parts and interactions). For this reason, mathematical and computational approaches are essential in GRN research.

Cellular behaviors have traditionally been characterized as being mediated through highly distinct processes (e.g., DNA replication) and pathways (e.g., the canonical WNT signaling pathway). However, because of widespread interactions among cellular processes and pathways, the use of unbiased, genome-wide technologies is essential to the discovery and characterization of GRNs.

In addition to the bedrock of 'classical' *cis*-regulatory analysis, GRN modeling today is buttressed by four cornerstones: (i) high-throughput technologies, (ii) integrative analysis of complementary data types, (iii) leveraging large-scale public datasets, (iv) computational modeling and analysis. This article reviews recent developments and discusses their implications for future research.

To maintain coherence and brevity, this review will focus on developments in human GRN modeling and analysis. Diverse new GRN modeling opportunities are also opening up in both well-studied and less-studied organisms. These and the complex GRNs underlying interactions between hosts and commensal or pathogenic organisms are beyond the scope of the present review.

Types and uses of human GRN modeling

A model is any representation of a system that can facilitate its analysis, communication, or documentation [3]. Modeling is at the heart of GRN research at multiple levels. At the most basic level, statistical models are at the heart of all high-throughput data analysis. For example, statistical models are commonly used to characterize DNA fragment length distribution as a first step towards the identification of transcription factor (TF) binding peaks in ChIP-seq (chromatin immunoprecipitation followed by high-throughput DNA sequencing) data.

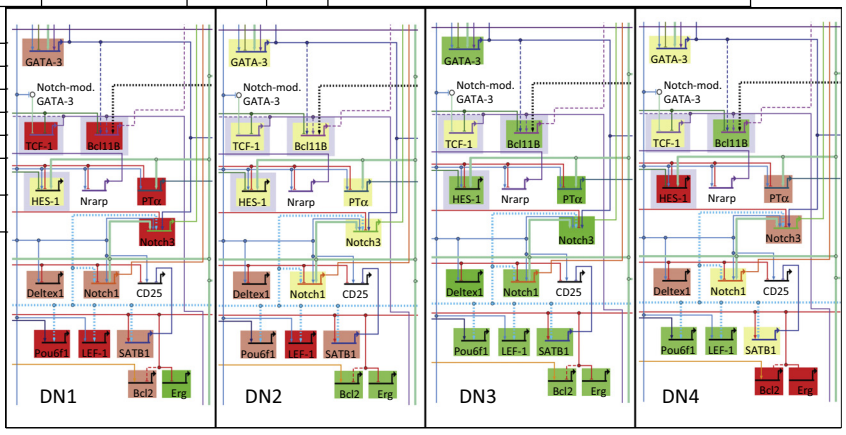
Given filtered data, methods such as network inference [4], guilt-by-association (e.g., through network or expression clustering; see Figures 2 and 3), and enrichment/over-representation analysis (e.g., to identify the impacted pathways or processes [5]) are used to organize genes and their products into broad-brush conceptual models. These models can then be refined and extended by integrating multiple data types each highlighting a different

Corresponding author: Bolouri, H. (HBolouri@fhcrc.org).
Keywords: gene regulatory networks; modeling; network biology; big data; computational biology; bioinformatics; systems biology.

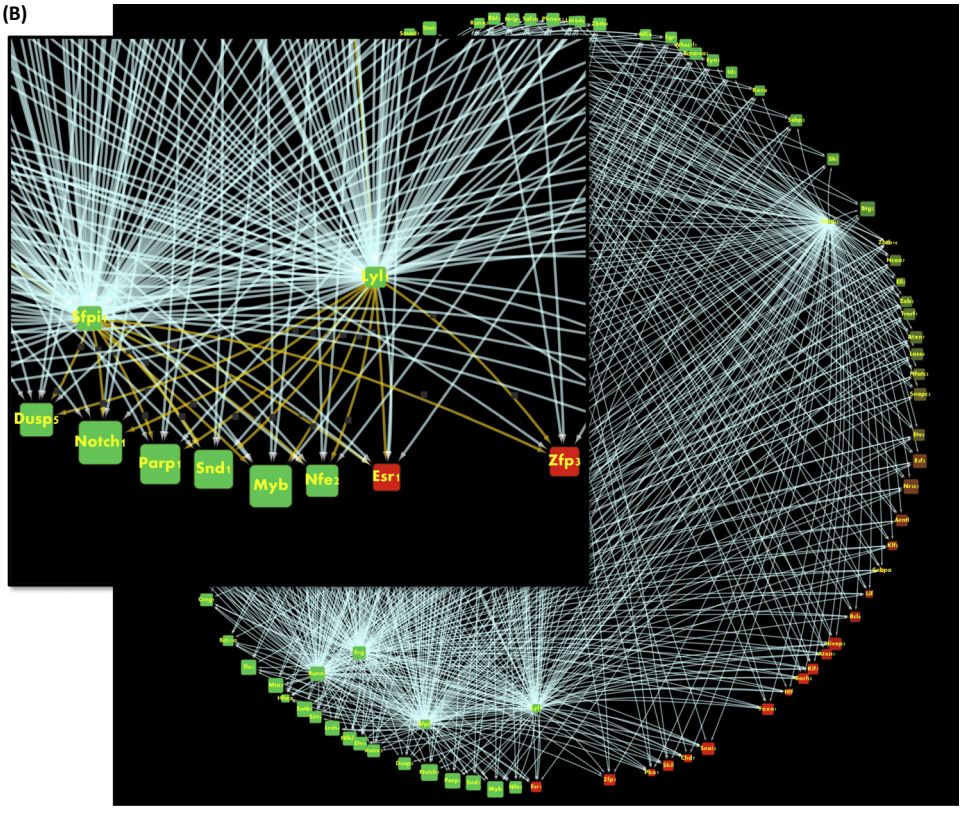
0168-9525/$ – see front matter
 http://dx.doi.org/10.1016/j.tig.2014.02.005

(A)

SourceId	TargetId	Source	Sign	Target	Evidence
32	171	Notch/CSL	REPRESSES	C/EBPa	Weak early effect? Stronger indirect effect? (Franco et al. 2006 shows effect on purified Thy-1+ cells; not seen in crude fraction in Taghon et al 2007)
265	100	DN2, DN3 Stage negative Regulator	REPRESSES	GATA-2	Tydell et al. 2007 J. Immunol.
264	163	DN3 Specific regulator	PROMOTES	Eva1	Tydell et al. 2007 J. Immunol.
265	161	DN2, DN3 Stage negative Regulator	REPRESSES	c-Kit	Taghon et al. 2006 Immunity; Yui & Rothenberg 2004 J. Immunol.
264	165	DN3 Specific regulator	PROMOTES	Notch1	Taghon et al. 2006 Immunity
192	82	Notch-mod. GATA-3	REPRESSES	TCF-1	Taghon et al 2007 Nat Immunol
32	117	Notch/CSL	PROMOTES	HEBAlt	Wang, D.et al.: Anderson, J Immunol. 20 PNAS.
32	160	Notch/CSL	PROMOTES	CD25	Taghon et al 2005 ; I. Maillardet al.: Pear, J Exp Med. 2006 Oct 2;203(10):2239-45.
32					
32					
32					
32					
32					
32					
32					



(B)



TRENDS in Genetics

Figure 1. Examples of gene regulatory network analysis, documentation, and visualization. **(A)** Part of BioTapestry visualization of a proposed early T cell specification gene regulatory network (GRN) (adapted from: http://www.its.caltech.edu/~tcellgrn/Oldnetwork.html). Each gene (symbol with a bent arrow) is represented as having a regulatory region (horizontal line) and a transcriptional output (arrow). A transcription factor (TF)–DNA binding interaction is depicted as an arrow incident on the regulatory region of a gene. Protein–protein interactions are depicted by circles with incident and output arrows. The background color of each gene indicates the fold-change in expression of the gene at a particular developmental stage. Snapshots of the network over four developmental stages are shown [double negative (DN) 1 to 4]. In the interactive viewer, clicking on a gene brings up a table showing the experimental data supporting the indicated regulatory interactions. **(B)** Cytoscape visualization of potential T cell specification gene regulatory interactions derived from ChIP-seq and gene expression data. Arrows represent regulatory interactions. Node colors and sizes represent gene expression levels at early and late developmental stages. The inset shows a zoomed-in view of the lower portion of the network. Using Cytoscape utilities, the user can quickly and easily identify a set of genes coregulated by Sfpi1 and Lyl1 (edge arrows highlighted in gold). This example network was derived during a 1.5 h introductory laboratory session by novice computational biology students (see http://www.bu.edu/computationalimmunology/summer-school/ for details).