PostDoc Journal
Vol. 1 No. 1, January 2013

Journal of Postdoctoral Research
www.postdocjournal.com

Estimation of Gene Regulatory Networks

Matthew N McCall[1]*
[1]Department of Biostatistics and Computational Biology,
University of Rochester Medical Center, Rochester, NY, USA

Complex gene regulatory networks, not individual genes, control cellular function. Genes and gene products act together to determine cellular phenotypes. Estimation of these networks is necessary to understand cellular mechanisms, detect differences in gene regulation between cell types, and predict cellular response to interventions. A plethora of algorithms have been developed to infer network structure from experimental data. Here we provide a general introduction to the estimation of gene regulatory networks and the classes of proposed algorithms.

Introduction

The central dogma of molecular biology states that information flows from DNA to RNA to proteins. While each of the cells of a complex organism contain the same DNA, different regions of DNA, called *genes*, can be transcribed into RNA and translated into proteins. This allows cells from different parts of the body to perform different tasks and an individual cell to alter its behavior in response to stimuli.

There are several techniques commonly used to measure gene expression. Many of these techniques rely on the *hybridization* of single-stranded DNA and RNA molecules, in which two complementary strands of DNA or RNA will bind to each other. This allows one to isolate a specific target from a mixture of DNA and/or RNA by designing a *probe* that is complementary to a specific region of the target molecule.

One of the first techniques used to measure gene expression was a *northern blot*, in which the expression of a single gene is assessed by examining the amount of RNA that binds to a labeled probe designed to perfectly complement a region of the target gene. The primary drawbacks of Northern blotting are that only a single RNA can be measured at a time and quantification of expression is based solely on visual examination of an image (Taniguchi et al., 2001). An alternative approach is *reverse transcription polymerase chain reaction (RT-PCR)*, in which a target RNA is reverse transcribed into cDNA, amplified (repeatedly replicated), and then measured. Quantification of expression is based on either the number of amplifications needed to achieve a predetermined threshold or via examination of the amplification curve (Bustin, 2005).

Recently, the measurement of gene expression has shifted to *high-throughput* methods, in which the expression of thousands of genes are measured simultaneously. The most widely used high-throughput technology to measure gene expression is the *DNA microarray.* A microarray typically contains thousands (sometimes millions) of probes each designed to hybridize to specific RNA molecules. Often millions of copies of each probe are present on a single microarray. Labelled RNA is then allowed to hybridize to the probes on the microarray, and the amount of each RNA present in the sample is quantified by the amount that hybridizes to the microarray. While microarrays provide a wealth of information, numerious statistical techniques were required to address various biases in this technology (Li and Wong, 2001; Irizarry et al., 2003; Wu et al., 2004; McCall et al., 2010). An alternative to DNA microarrays is *RNA-sequencing*, which determines the order of nucleotides within an RNA molecule. The application of sequencing techniques to measuring gene expression was extremely costly and time-consuming prior to recent advances in sequencing technology. These recent advances in sequencing techniques, termed *second-generation sequencing*, have made RNA-sequencing a viable method to measure high-throughput gene expression.

Over the past decade, numerous genes whose expression differs between conditions (tumor/normal,

*Email: mccallm@gmail.com

treated/untreated, etc.) have been reported. However, genes do not function in isolation, rather they act together in complex networks that drive cellular function. By considering the interactions between genes (and gene products), we gain a much deeper understanding of the underlying cellular mechanisms. For example, recent cancer research suggests that malignant transformation is the result of drastic changes in genetic networks critical to normal cellular function (Lloyd et al., 1997; McMurray et al., 2008; Xia and Land, 2007). Research aiming at the identification of intervention targets requires a detailed understanding of the gene regulatory networks present in a normal cell and the changes brought about by malignant transformation. Understanding of complex genetic networks has the potential to advance fields ranging from basic science research to clinical practice. Examination of cellular networks has provided insights in evolution (Isalan et al., 2008), metabolism (Ideker et al., 2001), DNA damage response (Bandyopadhyay et al., 2010), and cancer metastasis (Chuang et al., 2007).

In its most general form, a network consists of nodes and edges. What the nodes and edges signify determines the type of network. There are several types of cellular networks – e.g. metabolic networks, cell signaling networks, gene regulatory networks. In this article, we will restrict our focus to gene regulatory networks. Because genes encode proteins which are responsible for the vast majority of cellular function, networks that control which genes are expressed (transcribed into mRNA then translated into proteins) indirectly regulate the majority of cellular function. In a gene regulatory network, nodes represent genes and edges represent regulatory relationships.

The regulation of one gene by another is not carried out directly, rather the regulator encodes a protein that performs the regulation. This regulation can take many forms depending on the protein encoded. Proteins that bind to a specific DNA sequence and either increase or decrease transcription of a gene are called *transcription factors.* Transcription factors typically function by aiding or inhibiting the binding of RNA polymerase, the enzyme that transcribes DNA to RNA. Other proteins regulate gene expression without binding to DNA. These include proteins that are involved in chromatin remodeling, acetylation, or methylation. These result in changes in the accessibility of regions of DNA, and thereby changes in gene expression.

Input Data

Several types of data can be used to estimate gene regulatory networks. The type of input data often drives the choice of network to be estimated and the algorithm used to infer the network structure. The two most common types of data used to infer network structure are gene expression across biological replicates and gene expression after an experimental perturbation. For the former, the central idea is that genes that display similar expression profiles are coregulated. For the latter, one seeks to determine which genes respond to the perturbation of another gene.

From early knock-out experiments in Saccharomyces cerevisiae (Ideker et al., 2001; Winzeler et al., 1999) to RNAi and shRNA experiments in Caenorhabditis elegans, Drosophila melanogaster, mouse and human (Amit et al., 2009; Boutros and Ahringer, 2008; Fuchs and Boutros, 2006; Ivanova et al., 2006; Moffat and Sabatini, 2006), perturbation experiments have been used to better understand how cells function, to differentiate between diseased and normal cells, and to provide potential targets for intervention. While difficult and time-consuming, perturbation experiments have been shown to result in more reliable network reconstructions (Markowetz and Spang, 2003; Werhli et al., 2006; Zak et al., 2003). They also provide a straight-forward method to predict cellular response to intervention(s), allowing researchers to determine potential targets to produce a desired outcome. For example, if one could find an intervention target that results in apoptosis for gene networks present in a specific type of cancer and absent in normal cells, one could preferentially kill malignant cells.

In addition to be time-consuming and difficult to perform, perturbation experiments typically alter the gene regulatory network itself. For example, a persistent perturbation, in which a target gene is constitutively expressed, will result in a network in which any down-regulation of the target gene is masked. Algorithms should explicitly model this difference between the perturbed and the unperturbed network. Failing to do so may result in inaccurate modeling of the interactions present in the unperturbed cell.

Gene Connectivity Networks

Perhaps the simplest gene network is a *connectivity network* in which nodes are labeled by genes and an edge exists between two genes if there is an interaction between them. Such a network is said to be *directed* if the edges imply a causal relationship. In a directed network in which an edge goes from node i to node j, node i is called a *parent* of node j and node j is called an *offspring* of node i. While an undirected relationship simply states that two nodes are closely related, a directed relationship has a clear biological interpretation – the parent node regulates the offspring node. A directed network can be further classified as *cyclic* or *acyclic* depending on whether cycles exist (Figure 1).

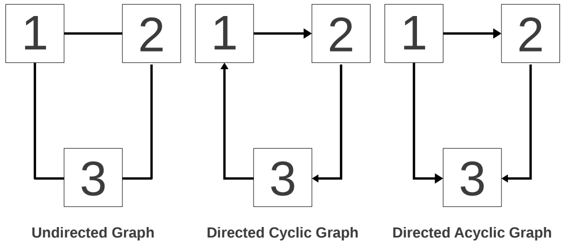

Figure 1: A three gene network showing the different types of connectivity networks.

A connectivity network can easily encoded in an *adjacency matrix*, $\mathbf{A}$, where

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathbf{E}\text{, the set of all edges} \\ 0 & \text{otherwise} \end{cases}$$

For an undirected network, $\mathbf{A}$ will be symmetric. For a directed network, one can let the columns represent parents and the rows represent children, such that $\mathbf{A}_{ij} = 1$ denotes that j is a parent of i.

A connectivity network can be generated from expression across biological replicates, in which case, the connectivity network is undirected with edges connecting genes with similar expression profiles. A connectivity network can also be created from perturbation experiments in which gene expression is assessed following experimental perturbation of a target gene. In this case, the connectivity network is directed with edges representing a change in the offspring gene following perturbation of the parent gene.

It is important to note that even a directed connectivity graph does not convey any information regarding the form of regulation between genes. For example, in the directed acyclic graph shown in Figure 1, it is unclear what functional form describes the dependence of gene 3 on genes 1 and 2. It is possible that expression of gene 3 requires the expression of both gene 1 and gene 2 or either could suffice. Moreover, the expression of gene 3 might be a continuous function of the expression of genes 1 and 2.

Coexpression Networks

Coexpression networks are generated from gene expression data by computing a measure of coexpression between pairs of genes. Genes that are highly coexpressed are assumed to be proximal in a gene regulatory network and genes that are lowly coexpressed distal. Typically, a threshold is used to determine whether an edge exists between two nodes – e.g. a correlation greater than 0.6 results in an edge. This typically results in an undirected network. Although coexpression cannot directly distinguish between direct and indirect interactions, it is typically assumed that directly interacting genes will have greater coexpression than indirectly interacting genes.

Coexpression can be measured in many different ways. The simplest and perhaps most widely used measure is correlation, which assesses the linear dependence between two genes. However, its inability to detect non-linear dependence makes it less suitable for more complex gene regulatory relationships. An alternative that is able to detect non-linear relationships is the *mutual information (MI)* (Butte and Kohane, 2000, 2003). Mutual informaiton is defined as follows:

$$MI(AB) = H(A) + H(B) - H(AB)$$

where $H(A)$ is the entropy of gene A, $H(B)$ is the entropy of gene B, and $H(AB)$ is the entropy of the joint distribution of genes A and B.

Whether one uses correlation or mutual information, to generate a network from coexpression data, one must decide on a threshold above which an edge will be inferred. Because coexpression is measured continuously, every gene pair will have a non-zero value. Furthermore, as previously mentioned, coexpression networks work on the assumption that direct interactions will have higher coexpression than indirect interactions; therefore, a network generated by a thresholding procedure will likely be dominated by direct interactions. For example, a network could be constructed by accepting edges associated with a correlation coefficient > 0.6 as in Zhou et al. (2002). A more rigorous approach would be to assess the signifance of coexpression measures via a permutation test in which one repeatedly randomly permutes the ordering of genes within replications and recalculates the coexpression measures. The permutation replicants form a null distribution that can be used to assess significance (Butte et al., 2000).

One of the primary challenges of this type of analysis is obtaining enough biological variability. Because the data consist of biological replicates, one must rely on subtle changes in gene expression to observe correlation between genes. Furthermore, the potential for technical variability to outweigh biological variability may result in a significant number of false positives. However, this type of network modeling has been successfully used to investigate functional relationships between gene expression and susceptibility to chemotherapeutic agents (Butte et al., 2000) and examine regulatory networks in human B cells (Basso et al., 2005).

Deterministic Networks

The simplest deterministic network is a Boolean Network in which nodes take on a value of either zero (unexpressed) or one (expressed). This model was originally introduced by Kauffman (1969). While discretizing gene expression is certainly a simplification, similar approaches have resulted in increased reproducibility and robustness when estimating both absolute and differential gene expression (McCall et al., 2011; Parmigiani et al., 2002; Scharpf et al., 2003; Zilliox and Irizarry, 2007) and have been used to correctly classify tumor types (Shmulevich and Zhang, 2002). Moreover, Boolean network models have been used to successfully model gene regulatory networks involved in the yeast cell-cycle (Li et al., 2004; Davidich and Bornholdt, 2008), cell-fate in Arabidopsis (Espinosa-Soto et al., 2004), and the mammalian cell-cycle (Faure et al., 2006).

Boolean networks are *dynamic*, meaning that they are governed by *transition functions* that take as input the current network state, i.e. which nodes are on (1) and which are off (0), and determine the subsequent states. They are also *deterministic*, meaning that the transition functions do not change. One can begin from any initial state and iteratively apply the transition functions until a state is repeated. Because the model is deterministic, once a state is repeated the network will continually cycle through the same sequence of states. This sequence of states is called an *attractor*, and an attractor that consists of only one state is called a *fixed point*. The set of initial states that lead to a given attractor is called its *basin of attraction*.

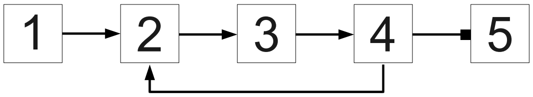

Figure 2: A five gene network where nodes represent genes and edges represent positive-regulation (arrows) or negative-regulation (squares). Note that this graph does not convey all of the necessary information – it is unclear whether expression of gene 2 requires expression of both genes 1 and 4 or whether either will suffice.

Considering the network in Figure 2, we can define a Boolean network as follows: gene 1 not regulated by any of the other 4 genes; gene 2 will be expressed if either gene 1 or gene 4 is expressed, otherwise it will remain unexpressed; gene 3 will be expressed if either gene 2 or gene 4 is expressed; gene 5 will be expressed if gene 4 is unexpressed. Note there are two primary attractors: (1) genes 1-4 unexpressed and gene 5 expressed, and (2) genes 2-4 expressed and genes 1 & 5 unexpressed. One can show that any initial state will result in one of these attractors, and once either of these states is reached, the network will remain in that state. In other words, both of these attractors are fixed points.

In addition to being a convienent modeling tool, there is growing evidence that attractors are present in cellular networks (Chang et al., 2008; Huang et al., 2005). One can view attractors as the determinants of cell type and functional state (e.g. liver cell during proliferation); cells are able to shift from one functional state attractor to another but not between cell type attractors. Such a view also provides potential insight into cancer, which can be viewed as cells spending too much time in a proliferation state and not entering the apoptosis state (Huang, 2001). This could potentially be the result of altered basins of attraction for these states

(Shmulevich and Aitchison, 2009). Basins of attraction can also be viewed as representing the robustness of a gene regulatory network. Transient perturbations, in which a target gene is altered but then allowed to revert to its former state, typically result in a transition within the original basin of attraction, meaning that the network will revert to the same attractor (Huang, 1999).

The fundamental Boolean network inference problem is finding transition functions that explain the observed data. The observed data are typically steady state measurements, representing expression once an attractor has been reached, although time course data are occasionally used. For genes whose expression is constant within the attractor, the steady state data will reflect this level of expression. For genes whose expression varies within the attractor, the data will represent a summary over the expression states. Therefore, one can view the inference problem as selecting transition functions constrained by the observed attractor summaries. If logical inconsistencies exist due to errors in the data, there may not exist such a set of transition functions. On the other hand, there may be multiple sets of transition functions that are able to explain the observed data. While previous work approached these challenges through "best-fit" solutions and parsimony (Liang et al., 1998; Akutsu et al., 1999; Ideker et al., 2000; Maki et al., 2001; Shmulevich et al., 2003; Lahdesmaki et al., 2003), recent work has proposed explicitly modeling this uncertainty via a posterior density (Almudevar et al., 2011).

Another approach is to use the Coefficient of Determination (Dougherty et al., 2000). The coefficient of determination measures the extent to which the expression of a given gene can be predicted by the expression of another set of genes. For a Boolean network, the coefficient of determination can be formulated as follows:

$$\theta_i = \frac{\varepsilon_i - \varepsilon}{\varepsilon_i}$$

where ε_i is the error from the best estimate of gene i in the absence of information from other genes and ε is the error from the optimal predictor of gene i based on all other genes. Note that $0 \leq \theta_i \leq 1$ with $\theta_i = 0$ when using information from other genes results in no improvement and increasing values of θ_i corresponding to greater reductions in error when using other genes to predict gene i. In practice, θ_i is unknown but can be estimated from training data; however, this process is computationally intensive for data in which a large number of genes are measured (Shmulevich and Dougherty, 2007).

One challenge in Boolean network inference is estimation of the initial state (Lee and Tzou, 2009). While there have been recent efforts to estimate absolute gene expression from microarray data (McCall et al., 2011) and RNA-sequencing (Mortazavi et al., 2008), estimates of differential gene expression are typically far more reliable because technical artifacts, such as probe-effects in microarray data, often cancel out. For this reason, it is often advantageous to assess gene expression from perturbation experiments relative to gene expression in unperturbed cells. For perturbation experiments in which gene expression has been assessed in unperturbed control cells, Boolean network models can be naturally extended to ternary network models by defining states as follows: under-expression (-1), baseline expression (0), and over-expression (1). This allows one to use estimates of differential expression to discretize gene expression (Kim et al., 2000).

Another criticism of Boolean networks is that the transition functions are typically applied to each node simultaneously. This is typically referred to as a *synchronous* network. Such a model may not be biologically plausible, since some genes may response far more quickly to their regulators than others. A simple solution to this criticism is to allow nodes to update asynchronously or to remove the notion of discrete time completely via a *continuous time boolean network* (Öktem et al., 2003). Finally, one can incorporate cellular dynamics via differential equations models to potentially better approximate actual cellular networks; however, these models are often very complex and require additional information, specifically kinetic constants.

Stochastic Networks

Unlike deterministic networks, stochastic networks view the network structure as random in nature. The majority of deterministic networks can be modified to add a random component thereby making them stochastic. For example, a Boolean network can be modified such that at each iteration, one of several transition functions is chosen probabilisticly for a given node.

The most widely used stochastic network is a

Bayesian network. A Bayesian network is defined by a set of nodes which are viewed as random variables and a set of directed edges which are specified by conditional probabilities. The values of the nodes can be either continuous or discrete depending on the form of the conditional probabilities. However, a Bayesian network must be acyclic. This is perhaps the greatest drawback to the application of Bayesian networks to cellular regulation because cells contain numerous feedback loops.

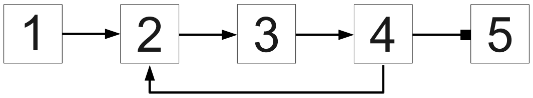

Figure 3: A five gene Bayesian network where nodes represent genes and edges represent regulatory relationships.

For a Bayesian network such as the one show in Figure 3, the graph contains relatively little information – only which genes regulate which other genes. The network itself is specified by the conditional probabilities that describe the regulatory relationships. For example, the network shown in Figure 3 could be described by the following conditional probabilities:

$$\begin{aligned}
&P(G_1 = 1) = 0.2\\
&P(G_2 = 1 \mid G_1, G_4) = 0.7 * \mathbf{1}(G_1 = 1, G_4 = 1)\\
&P(G_3 = 1 \mid G_1, G_5) = 0.3 * \mathbf{1}(G_1 = 1, G_5 = 0)\\
&P(G_4 = 1) = 0.9\\
&P(G_5 = 1 \mid G_4) = 0.1 + 0.6 * \mathbf{1}(G_4 = 1)
\end{aligned}$$

Like the Boolean network described above, the genes in this network only take on values of zero (unexpressed) or one (expressed), but unlike the Boolean network, the expression of the genes in this network are stochastic. For example, expression of gene 4 increased the probability that gene 5 is expressed (by 0.6), but it does not guarentee that gene 5 will be expressed.

One method to create a cyclic Bayesian Network is to include a time component such that the conditional probabilities governing each node are allowed to change over time. This results in a *Dynamic Bayesian Network* (Murphy et al., 1999), which allows the network to incorporate regulatory feedback via cycles that exist over time.

A stochastic version of a Boolean Network, called a *Probabilistic Boolean Network*, was originally proposed to deal with uncertainty in inferring Boolean Networks from relatively noisy data and small sample sizes (Shmulevich and Dougherty, 2007). To address the potential error in inferring network structure from the observed data, the transition functions are chosen probabilisticly. That is, a Probabilistic Boolean Network extends a Boolean Network by allowing each node to be governed by more than one transition function. When updating a node, the transition function used is determined randomly from a set of possible transition functions. This random selection can be weighted toward more likely functions or can be reweighted based upon the functions chosen for other nodes. If one uses coefficients of determination to select transition functions, one can also use the coefficient of determination to weight the probability of each transition function being employed (Shmulevich et al., 2002).

Stochastic networks, particularly Bayesian networks, have seen fairly extensive application in genomic biology. For example, they have been used to investigate Her2 signaling in breast cancer (Bose et al., 2006) and as part of an integrated approach to discover genes driving cancer (Akavia et al., 2010).

Network Inference

Many network inference algorithms function by *scoring* potential models and selecting models with a better score. Such a score is a function of the model and the data with models that better reflect the observed data scoring better. Because one is often comparing models of different complexitity, it is standard to include a penalty for added complexity – the score then contains a component representing how well the model matches the data and a penalty for model complexity. This controls over-fitting because for any given model, it is almost always possible to find a more complex model that performs as well or better.

One major challenge in inferring gene regulatory networks is that the number of genes (nodes) typically far exceeds the number of samples – e.g. in genomic data one typically measures tens of thousands of genes but with at most hundreds of samples. One approach to reduce the model space is to restrict the complexity of the network – for example, by limiting the maximum number of regulators of any gene or searching for a small set of genes that explain the majority of the observed expression (Pe'er et al., 2002). An alternative approach to reduce the model space is to group genes into gene sets that are collectively regulated, reducing the number of nodes (Segal et al., 2003).

Even after employing algorithms to reduce the model space, the number of possible networks is often still too large to score every possible network in a reasonable amount of time. Moreover, the set of networks compatible with the data can be very large and may contain networks with very different structure. Nonetheless, a single network that optimizes some criterion, e.g. minimizing the number of differences between observed and predicted attractors or maximizing parsimony (Ideker et al., 2000; Lahdesmaki et al., 2003), is typically reported. However, a single network that completely explains the observed data often does not exist – errors in the input data producing logical inconsistencies may preclude any network perfectly fitting the observed data. On the other hand, when solutions do exist they will rarely be unique. In other words, in the absence of logical inconsistencies, there are often multiple networks that explain the observed data. Therefore, the model space is typically explored using complex search algorithms to construct a posterior distribution on the model space. These methods often rely on simulated sampling, typically Markov Chain Monte Carcl (MCMC) samplers. The feasibility of this approach has been demonstrated for Bayesian networks (Friedman and Koller, 2003) and Boolean networks (Almudevar et al., 2011).

Conclusions

Over the past decade, advances in the estimation of gene regulatory networks have lead to an increased understanding of cellular regulation. In particular, the increased use of targeted gene perturbation experiments promises richer data regarding the cellular processes involved in a wide variety of diseases and the potential to design targeted interventions. However, additional methodological advances are still needed as current network estimation algorithms are unable to adequately reconstruct gene networks from gene expression data alone (Marbach et al., 2010). In particular, many network estimation algorithms struggle to accurately detect complex multi-gene regulatory relationships.

One possible path forward is to supplement gene expression data with additional information. This additional information can be measurements of transcription factor binding via Chip-chip or Chip-seq (Hartemink et al., 2002), protein-protein interactions (Imoto et al., 2004), or even literature mining (Haibe-Kains et al., 2012). Combining these additional data sources with gene expression data, ideally following targeted perturbations, may allow researchers to uncover the complex networks that govern cellular processes.

Competing Interests

The author declares that he has no competing interests.

Acknowledgements and Funding

The work of MNM was funded by National Institutes of Health (T32CA009363) and a grant from the Edelman-Gardner Foundation.

References

U.D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H.C. Causton, P. Pochanard, E. Mozes, L.A. Garraway, and D. Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017, 2010.

T. Akutsu, S. Miyano, S. Kuhara, et al. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing*, volume 4, pages 17–28. World Scientific Maui, Hawaii, 1999.

A. Almudevar, M.N. McCall, H. McMurray, and H. Land. Fitting Boolean Networks from Steady State Perturbation Data. *Statistical applications in genetics and molecular biology*, 10(1):47, 2011.

I. Amit, M. Garber, N. Chevrier, A.P. Leite, Y. Donner, T. Eisenhaure, M. Guttman, J.K. Grenier, W. Li, O. Zuk, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, 326(5950):257, 2009. ISSN 0036-8075.

S. Bandyopadhyay, M. Mehta, D. Kuo, M.K. Sung, R. Chuang, E.J. Jaehnig, B. Bodenmiller, K. Licon, W. Copeland, M. Shales, et al. Rewiring of genetic networks in response to dna damage. *Science*, 330 (6009):1385, 2010.

K. Basso, A.A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382–390, 2005.

R. Bose, H. Molina, A.S. Patterson, J.K. Bitok, B. Periaswamy, J.S. Bader, A. Pandey, and P.A. Cole. Phosphoproteomic analysis of her2/neu signaling and inhibition. *Proceedings of the National Academy of Sciences*, 103(26):9773–9778, 2006.

M. Boutros and J. Ahringer. The art and design of genetic screens: RNA interference. *Nature Reviews Genetics*, 9(7):554–566, 2008. ISSN 1471-0056.

S.A. Bustin. Real-time, fluorescence-based quantitative pcr: a snapshot of current procedures and preferences. *Expert review of molecular diagnostics*, 5(4):493–498, 2005.

A. Butte and I. Kohane. Relevance networks: a first step toward finding genetic regulatory networks within microarray data. *The Analysis of Gene Expression Data*, pages 428–446, 2003.

A.J. Butte and I.S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac Symp Biocomput*, volume 5, pages 418–429, 2000.

A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186, 2000.

H.H. Chang, M. Hemberg, M. Barahona, D.E. Ingber, and S. Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–547, 2008.

H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(1), 2007.

M.I. Davidich and S. Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PLoS One*, 3(2):e1672, 2008.

E.R. Dougherty, S. Kim, and Y. Chen. Coefficient of determination in nonlinear signal processing. *Signal Processing*, 80(10):2219–2235, 2000.

C. Espinosa-Soto, P. Padilla-Longoria, and E.R. Alvarez-Buylla. A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *The Plant Cell Online*, 16 (11):2923, 2004.

A. Faure, A. Naldi, C. Chaouiya, and D. Thieffry. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124, 2006.

N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine learning*, 50(1): 95–125, 2003. ISSN 0885-6125.

F. Fuchs and M. Boutros. Cellular phenotyping by RNAi. *Briefings in functional genomics & proteomics*, 5(1):52, 2006. ISSN 2041-2649.

B. Haibe-Kains, C. Olsen, A. Djebbari, G. Bontempi, M. Correll, C. Bouton, and J. Quackenbush. Predictive networks: a flexible, open source, web application for integration and analysis of human gene networks. *Nucleic Acids Research*, 40(D1):D866–D875, 2012.

A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, and R.A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'02)*, pages 437–449, 2002.

S. Huang. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *Journal of Molecular Medicine*, 77(6):469–480, 1999.

S. Huang. Genomics, complexity and drug discovery: insights from boolean network models of cellular regulation. *Pharmacogenomics*, 2(3):203–222, 2001.

S. Huang, G. Eichler, Y. Bar-Yam, and D.E. Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical review letters*, 94(12):128701, 2005.

T. Ideker, V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929, 2001. ISSN 0036-8075.

T.E. Ideker, V. Thorsson, and R.M. Karp. Discovery of regulatory interactions through perturbation: inference and experimental design. In *Pacific Symposium on Biocomputing*, volume 5, pages 302–313. Citeseer, 2000.

S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *Journal of Bioinformatics and Computational Biology*, 2(01):77–98, 2004.

R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, UWE Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4 (2):249, 2003.

M. Isalan, C. Lemerle, K. Michalodimitrakis, P. Beltrao, C. Horn, E. Raineri, M. Garriga-Canut, and L. Serrano. Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, 452(7189):840, 2008.

N. Ivanova, R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. DeCoste, X. Schafer, Y. Lun, and I.R. Lemischka. Dissecting self-renewal in stem cells with RNA interference. *Nature*, 442(7102):533–538, 2006. ISSN 0028-0836.

S. Kauffman. Homeostasis and differentiation in random genetic control networks. *Nature*, 224:177–178, 1969.

S. Kim, E.R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J.M. Trent, M. Bittner, et al. Multivariate measurement of gene expression relationships. *Genomics*, 67(2):201–209, 2000.

H. Lahdesmaki, I. Shmulevich, and O. Yli-Harja. On learning gene regulatory networks under the Boolean network model. *Machine Learning*, 52(1):147–167, 2003. ISSN 0885-6125.

W.P. Lee and W.S. Tzou. Computational methods for discovering gene networks from expression data. *Briefings in bioinformatics*, 10(4):408, 2009.

C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *PNAS*, 98(1):31, 2001.

F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4781, 2004.

S. Liang, S. Fuhrman, R. Somogyi, et al. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific symposium on biocomputing*, volume 3, pages 18–29, 1998.

A. C. Lloyd, F. Obermuller, S. Staddon, C. F. Barth, M. McMahon, and H. Land. Cooperating oncogenes converge to regulate cyclin/cdk complexes. *Genes Dev*, 11:663–677, 1997.

Y. Maki, D. Tominaga, M. Okamoto, S. Watanabe, Y. Eguchi, et al. Development of a system for the inference of large scale genetic networks. In *Pac. Symp. Biocomput*, volume 6, pages 446–458. Citeseer, 2001.

D. Marbach, R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, 107(14):6286–6291, 2010.

F. Markowetz and R. Spang. Evaluating the effect of perturbations in reconstructing network topologies. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 2003.

M.N. McCall, B.M. Bolstad, and R.A. Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11 (2):242, 2010.

M.N. McCall, K. Uppal, H.A. Jaffee, M.J. Zilliox, and R.A. Irizarry. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(suppl 1):D1011, 2011. ISSN 0305-1048.

H.R. McMurray, E.R. Sampson, G. Compitello, C. Kinsey, L. Newman, B. Smith, S.R. Chen, L. Klebanov, P. Salzman, A. Yakovlev, and Land H. Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype. *Nature*, 453(7198):1112–1116, 2008.

J. Moffat and D.M. Sabatini. Building mammalian signalling pathways with RNAi screens. *Nature Reviews Molecular Cell Biology*, 7(3):177–187, 2006. ISSN 1471-0072.

A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

K. Murphy, S. Mian, et al. Modelling gene expression data using dynamic bayesian networks. Technical report, Technical report, Computer Science Division, University of California, Berkeley, CA, 1999.

H. Öktem, R. Pearson, and K. Egiazarian. An adjustable aperiodic model class of genomic interactions using continuous time boolean networks (boolean delay equations). *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(4):1167–1174, 2003.

G. Parmigiani, E.S. Garrett, R. Anbazhagan, and E. Gabrielson. A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):717–736, 2002.

D. Pe'er, A. Regev, and A. Tanay. Minreg: inferring an active regulator set. *Bioinformatics*, 18(suppl 1): S258–S267, 2002.

R. Scharpf, E.S. Garrett, J. Hu, and G. Parmigiani. Statistical modeling and visualization of molecular profiles in cancer. *BioTechniques*, 34:S22–S29, 2003.

E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166–176, 2003.

I. Shmulevich and J.D. Aitchison. Deterministic and stochastic models of genetic regulatory networks. *Methods in enzymology*, 467:335–356, 2009.

I. Shmulevich and E.R. Dougherty. *Genomic Signal Processing*. Princeton University Press, 2007.

I. Shmulevich and W. Zhang. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*, 18(4):555, 2002.

I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261, 2002. ISSN 1367-4803.

I. Shmulevich, A. Saarinen, O. Yli-Harja, and J. Astola. Inference of genetic regulatory networks via best-fit extensions. *Computational and Statistical Approaches to Genomics*, pages 197–210, 2003.

M. Taniguchi, K. Miura, H. Iwao, S. Yamanaka, et al. Quantitative assessment of dna microarrays–comparison with northern blot analyses. *Genomics*, 71(1):34, 2001.

A.V. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, 22(20):2523, 2006. ISSN 1367-4803.

E.A. Winzeler, D.D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J.D. Boeke, H. Bussey, et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901, 1999. ISSN 0036-8075.

Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

M. Xia and H. Land. Tumor suppressor p53 restricts ras stimulation of rhoa and cancer cell motility. *Nature Structural & Molecular Biology*, 14(3):215–223, 2007.

D.E. Zak, G.E. Gonye, J.S. Schwaber, and F.J. Doyle. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. *Genome research*, 13(11):2396, 2003. ISSN 1088-9051.

X. Zhou, M.C.J. Kao, and W.H. Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences*, 99(20):12783–12788, 2002.

M.J. Zilliox and R.A. Irizarry. A gene expression bar code for microarray data. *Nature methods*, 4(11):911–913, 2007.