Process Pathway Inference via Time Series Analysis

by C.H. Wiggins and I. Nemenman

ABSTRACT—Motivated by recent experimental developments in functional genomics, we construct and test a numerical technique for inferring process pathways, in which one process calls another process, from time series data. We validate using a case in which data are readily available and we formulate an extension, appropriate for genetic regulatory networks, which exploits Bayesian inference and in which the present-day undersampling is compensated for by prior understanding of genetic regulation.

KEY WORDS—Genomics, pathways, gene expression regulation, Bayesian statistics, auto-regressive models

Motivation

The last decade has witnessed stunning advances in experimental biology, particularly in the fields of neuroscience and genomics, which have made possible "data-driven" biological investigations. As examples, the quantitative revolution of genomics has provided terabytes of transcriptome data; and neuroscientists routinely record for hours or even days from multiple neurons simultaneously. This transformation stands as a challenge to theorists who hope to advance understanding by making a connection between experiment and first principles models.[†]

In genomics, for example, we are presented with the expression levels of thousands of genes, but our ability to model is limited not only quantitatively, in that there are myriad unknown rate constants and binding parameters, but qualitatively, in that a sizable fraction of proteins and genes remain of uncharacterized function.[1] Similarly, in neuroscience, we can model patches of cellular membranes, synapses, and (at least electro-physiologically) entire cells.[2] However, this modeling hinges on numerous unknown parameters, and even if we can perform massive computations involved in the study of even rather small biological neural networks, the sensitivity to these parameters still makes the whole approach intractable. The astronomical amounts of experimental data are troubling

computationally, but even more immediate problems are the lack of reductive descriptions of the underlying phenomena and undersampling—an inability of the data to determine the (slightly smaller) astronomical number of important microscopic parameters appearing in theoretical models.

Presented with such an imbalance, it is important to distinguish among the possible questions we can ask as well as the possible tools at our disposal for answering them. That is, we can ask *what* the system is doing (a non-trivial question when the language for discovery is an astronomical number of unorganized data) before we ask *how* it is doing what it does. The latter involves building models of some microscopic fidelity. The former may be answered without reference to microscopics by a model-independent, data-driven phenomenological approach.

A useful historical analogy is that of particle physics of the late 1950s, in which an explosion of data from accelerators was equally daunting and similarly irreducible. At that time physicists were not yet asking the *how* questions (cross-sections, isospin multiplets, etc.) but were instead carefully, statistically, inferring the presence or absence of features in the data; for example, exploiting prior knowledge of quantum mechanics to constrain reasonable shapes for peaks in the data (hallmarks of newly discovered particles—the "resonances").

In this analogy, neuroscience is still dealing with the existence of peaks. Indeed, only recently (see Rieke et al.[3] and further works by the same authors) it has become clear that precise timings of single spikes are very important for understanding the neural code. This is a basic, objective, model-independent observation, and it is not surprising that Shannon's information theory, which was specifically designed with these types of questions in mind, turned out to be extremely useful.

In genomics, however, the quantities of interest are easier to identify; gene expressions are largely governed by the underlying regulation networks. Now is the time to attempt to infer these networks—still the *what* question—which corresponds to inferring the peaks in the data. This is a requisite step before classifying the possible networks and explaining the classification rules; an answer to *how* the system does what it does. Trying to answer *how* before carefully exploring *what* might ultimately produce many epi-manipulations of the data, but little significant understanding.

In other words, presented with data describing natural phenomena, we should form a phenomenology of experimental results, then inferences from the data in light of this phenomenology, and finally microscopic models. Genomics is currently at the penultimate step and, armed with careful *informatics*, here meaning the incorporation of data with prior knowledge in the absence of detailed models, we hope to

---

[†]The mathematization of such models are referred to below as the "microscopic equations"; consider, for example, those of fluid dynamics which govern, yet certainly fail to encapsulate, such phenomena as turbulence and the tumbling of a falling leaf.

C.H. Wiggins (chris.wiggins@columbia.edu) is an Assistant Professor, Department of Applied Physics and Applied Mathematics, and Center for Computational Biology and Bioinformatics (C2B2), Columbia University, NY 10027, and Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106. I. Nemenman is a Post-Doctoral Research Scientist, Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106.

reduce the data into a representation which allows description, prediction, and ultimately control.

An example of data reduction convenient for representation is cataloging of the regulatory networks. However, such cataloging is not a model-independent task; at the very least, our microscopic model includes the existence of the networks. Furthermore, even if we are only interested in a network's connection diagram and do not care much about the exact details of the connections, identification of the network still involves determination of many parameters. Thus it is not clear that information–theoretic approaches will be of great use. However, it is plausible that our intuition of how the underlying microscopic dynamics translates into macroscopic probabilistic models may play a big role. The main purpose of this paper is to show that this intuition, appropriately mathematized in a principled way as the *a priori* knowledge—the priors in Bayesian statistics—may be self-consistently incorporated into a macroscopic probabilistic model of process pathways without detailed, sophisticated modeling of microscopic dynamics. We will first show this on a simple synthetic example, and then suggest some extensions of the ideas with an eye towards genetic regulatory networks.

Functional Genomics

As mentioned, the motivating problem here is time series informatics applied to functional genomics.[†] We therefore briefly review genetics and characterize the relevant experiments, the data from which will be used in inferring the underlying connectivities and possibly control.

A Brief Review of Genetics

The central goal of functional genomics is the understanding of the interactions among distinct parts of the genome: the *genes*. Each gene consists of long words composed of thousands of coding base pairs of DNA which are then transcribed into mRNA, which is then translated into protein. Many of these proteins, called *transcription factors*, then regulate the rate of production of mRNA transcribed either by their own genes or other genes. The working of all the genes thus forms a *genetic regulatory network*, and may be thought of as a dynamical system. Inputs include elements of the physical world which affect the activity of the transcription factors, and outputs may be considered as the concentrations of the translated proteins or, at a deeper level, the transcribed mRNA. While the proteins are ultimately responsible for cellular function, the mRNA are more easily experimentally measured via *DNA microarrays*.

A Brief Review of DNA Microarrays

Only recently has it become possible to probe the expression of a number of genes comparable to the total number of genes in the entire genome of an organism via microarrays of nucleic acids, commonly known as "DNA chips". The most common application of such a chip is to monitor simultaneously the expression of thousands of genes by detecting hybridization of nucleic acid originating from a biological sample to target nucleic acids lying on the chip. We

can then probe, for example, the differences in gene expression between cancerous and non-cancerous cells of the same specialization,[4–10] or the expression of different genes as a function of the phase of the cell cycle,[11,12] or of the response of cells to chemical or physical perturbation. The two latter types of experiments produce time series of gene expressions, and they will be the focal point of our discussion from now on.

The first DNA microarrays were made by Affymetrix in the early 1990s.[13–15] In this technique, DNA oligonucleotides are attached to a surface in a specific spatial pattern, directed by optically activated chemical synthesis. One can build an arbitrary oligonucleotide sequence in a small area (approximately 20 μm per target) on the surface. However, the initial setup cost of creating the chip makes the technique infeasible for any application for which less than several hundred masks will be created (and sold). Individual researchers are completely without flexibility to change the chip to fit a particular area of investigation.

Functional genomics further benefited from a second technology in 1996, when Pat Brown's lab at Stanford introduced the spot chip.[16,17] This highly customizable technique exploits robotic deposition of drops on a microscope slide. The automation makes creating new and different slides a simple operation. Moreover, typically 120 slides at a time can be created. Individual researchers can thus design custom experiments, placing genes at locations or redundancies of their choosing. The gene fragments used in the spotting technique are hundreds of base pairs in length, and therefore less sensitive to single base pair mismatches.

Methods

Chemical Network Reconstruction

Genomic data are certainly not the first dynamic data for which reverse engineering of the interaction network has been attempted. A similar problem has been faced historically in chemistry, in which one would like to infer the underlying reactions responsible for observed data. Such reaction networks typically are sparse, that is, the typical connectivity is far less than the total number of chemical species. This is also true in genomics, where one gene typically interacts with no more than a few dozen others.[18]

To highlight the parallels, we may state the question as follows. Armed with sufficient temporal data taken from a number of interacting reagents (here, chemicals), is it possible to infer the circuit diagram? One possible strategy was proposed in 1995,[19] tested first on simulated data, and later on the glycolytic pathway,[20] and recently refined in light of ideas from information theory.[21] However, this strategy has yet to be successfully applied to any reactions which were not known by the authors beforehand, nor subjected to a "blind" test, as in the annual CASP (Critical Assessment of techniques for protein Structure Prediction) test among the protein folding community (see http://predictioncenter.llnl.gov). In addition, unlike in chemical kinetics, where the data are produced by moderately nonlinear and rapidly interrogated dynamical systems, genomic data sets are highly undersampled and are more like a set of fuzzy logic gates, or leaky Boolean circuits. Thus, successful application of techniques inspired by chemical networks to genomic data is, at best, doubtful.

We may nonetheless attempt reverse engineering of regulatory networks with a similar philosophy, in that, rather

than trying to fit to a precise microscopic model, we attempt to parametrize a minimal phenomenological model and infer macroscopic parameters from it.

Synthetic Network Reconstruction

In order to test any new attempt to infer connectivity from dynamics, it is useful to study a system which is qualitatively similar, e.g., which demonstrates degrees of freedom which turn on and off other degrees of freedom in a near-complete or "fuzzy logic gate" way and for which connectivity is sparse; yet for which data are readily available and obvious to interpret. To that end, we collected data from a multiuser UNIX machine, recording the relative CPU usage of all processes (the analog of mRNA concentration), user ID, and process name, as a function of time. This can be done in an automated way via `bash` script ((GNU) Bourne–Again SHell) and the results analyzed via MATLAB.

A typical time course, automatically labeled via MATLAB, is shown in Fig. 1 for a particular (anonymous) user at a large department of applied mathematics.

One axis is the job "number", ordered by frequency of occurrence over all users; the other axis is time (roughly in seconds). The height (and color) indicates CPU usage.

Modeling Synthetic Data

We begin with the minimal probabilistic model of process pathways, in which the strength of each process at subsequent time steps is linearly determined by the strength of all current processes. Similar linear models have been used with some success in understanding genomic data, including clustering via dynamics.[22] The most general model in discrete time is the $AR(p)$ model, in which we include the possibility that the state now is a function of the p previous observations of the system. We do not yet include the possibility of hidden degrees of freedom. Mathematically, we may pose the model as

$$\begin{aligned} \mathbf{g}_t &= \mathbf{w}_0 + \mathbb{M}_1\mathbf{g}_{t-1} + \mathbb{M}_2\mathbf{g}_{t-2} + \dots \\ &\quad + \mathbb{M}_p\mathbf{g}_{t-p} + \boldsymbol{\xi}_t \end{aligned} \tag{1}$$

$$\langle \xi_{i,t}\xi_{j,t'} \rangle = C^{ij}\delta_{t,t'}, \tag{2}$$

where the degree of freedom at observation t is $\mathbf{g}_t$, the transition matrices are $\mathbb{M}_j$, and the noise correlation $\mathbb{C}$ is as yet undetermined.

We may fit for the most probable transition matrices $\mathbb{M}_j$ as well as the offset $\mathbf{w}_0$ and the noise correlation matrix $\mathbb{C}$, using, for example, the standard Schwartz's Bayesian Information Criterion[23] to determine the most likely value of p. (See the section on Bayesian statistics for a brief discussion of this model selection technique.) Excellent numerical techniques and general purpose libraries have been designed for solving this problem.[24]

Note that we *do not claim* the actual interactions among the processes are linear, as eq (1) seems to imply. Indeed, as stated above, the exact values of the transition matrices $\mathbb{M}_j$ are of very little interest to us, and we are only interested in the topological features of the network. It is reasonable to expect that the absence of a connection between two processes will be fitted well by a zero in the corresponding transition matrix element, while the presence of a connection of any type will result in its non-zero value. The mismatch between the linear form of eq (1) and the actual dynamics will manifest itself

in a large variance C^{ij}. However, if we are not interested in the exact values of $\mathbb{M}_j$, this should not adversely effect our determination of connections.

Results

Fitting the observed CPU usages to the transition state model, we find that the most probable p value is 1, indicating a lack of inertia in the system; the resulting transition matrix $\mathbb{M}_1$ is plotted in Fig. 2. The noise covariance matrix was quite small, despite the naïveté of the model.

Causal connections between jobs are labeled with "→", e.g., "emacs→latex" or "emacs drives latex". We highlight several remarkable features:

1. Processes familiar to anyone who has used the typesetting software LATEX will be readily apparent: one edits a file (e.g., in `emacs`, `xemacs`, or `vi`), then compiles with `latex`, and views the result in `xdvi` and finally ghostscript ("`gs`"). Similarly one observes `emacs` drives `latex`, `latex` drives `gs`, etc.

2. Note that the matrix is not symmetric; one axis describes processes which "activate" other processes; the second describes which process is acted on. For example, `latex` drives ghostscript but ghostscript does not drive `latex`.

3. The transition matrix shows "upregulation" as well as "downregulation": some processes discourage other processes at later times.

4. Diagonal elements have not been labeled as they simply describe the likelihood the process will continue on to the next time step.

5. Note also how the transition matrix correctly infers the highly sparse connectivity of these disparate jobs. The vast majority of elements are 0, as they should be, since processes are not influenced by those called by other users.

For comparison, in Fig. 3 we also show an example of a transition matrix when the CPU usages are replaced with randomly generated data. Any reasonable structure is absent here. These results are in accord with our intuition that the proposed probabilistic model for data reduction, eq (1), although an incomplete description, still leads to a reasonable reconstruction of network connectivity.

Modeling Gene Regulation

In the example above, as mentioned, data are plentiful. In genomic time series, data are scarce. However, the above exercise is designed to test a phenomenological model into which we can incorporate additional knowledge about genetic regulation. This goal, constraining possible models by compensating for sparse data with prior knowledge, is mathematized via Bayesian analysis.

We emphasize that we are *not* using Bayesian analysis to attempt to fit for the innumerable unknowns in a microscopic model, e.g., a chemical kinetics model of transcriptional regulation. We are not interested in these parameters but in couplings and network topology. We instead augment
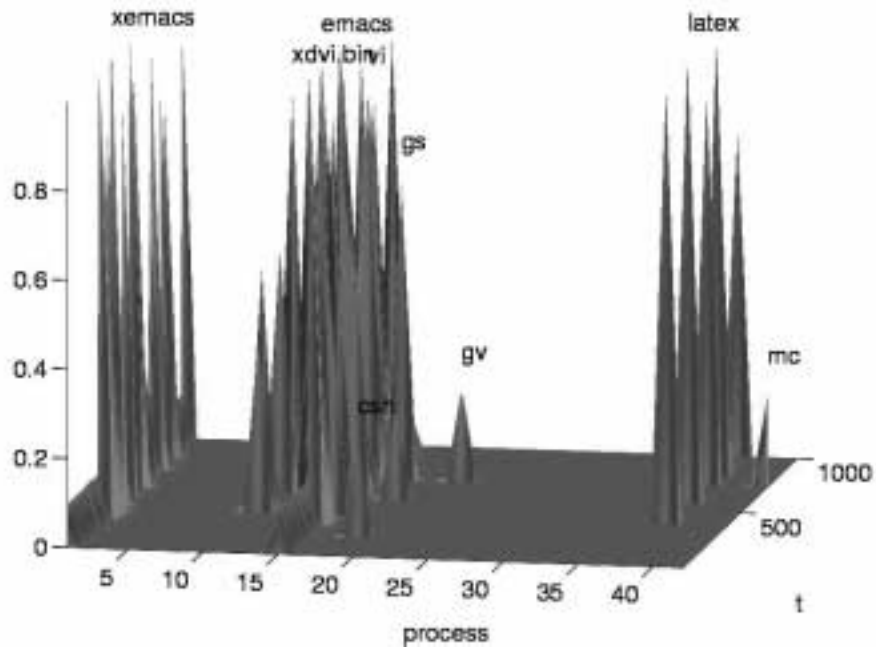
Fig. 1—CPU percentage for all the processes called by a particular user during the observation (approximately 10^3 seconds). Processes are numbered according to their relative frequency over *all* users during the observation time.
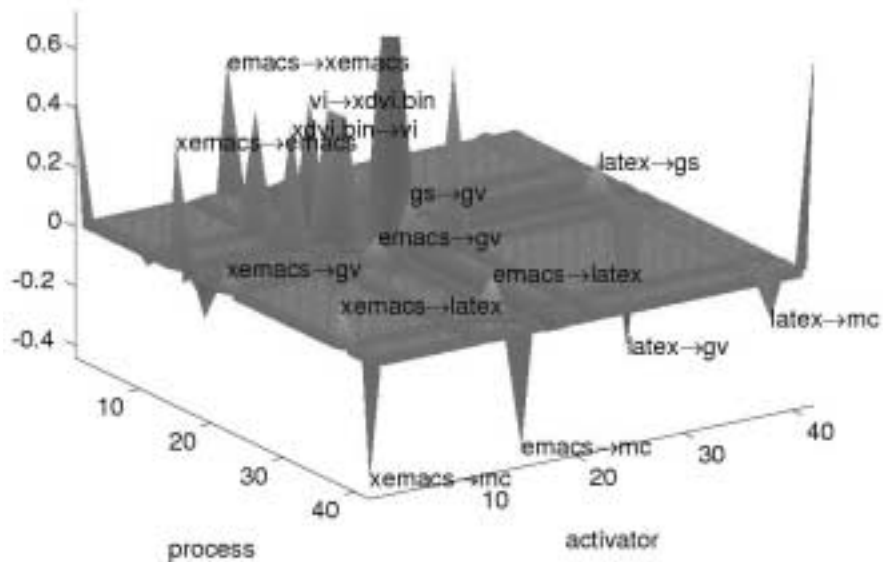

Fig. 2—The transition state matrix $\mathbb{M}_1$ resulting from the data in Fig. 1

the successful phenomenological or "macroscopic" model above with prior knowledge about transcriptional regulation.

Bayesian Statistics

A brief summary of Bayesian statistics is in order. We refer the interested reader to standard textbooks (cf. Press[25]) for discussions of philosophical implications of Bayesian statistics, as well as standard statistical properties of Bayesian estimators. We focus below only on the relevant features.

Bayes rule itself,

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)} , \qquad (3)$$

is merely a rewriting of the rules of joint probabilities. The connection with interpretation of experiments is made by identifying b as the model, or specifically the vector of parameters θ of the model; and a as the data D. Then
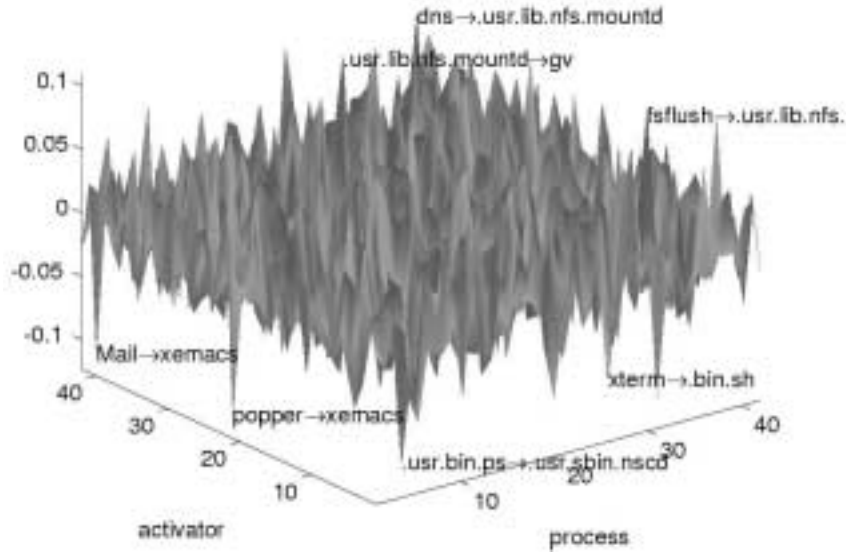
Fig. 3—The transition state matrix $\mathbb{M}_1$ resulting from random data

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}\,, \qquad (4)$$

$$P(D) = \int d\theta\, P(\theta, D) = \int d\theta\, P(D|\theta)P(\theta). \qquad (5)$$

In eq (4), the left-hand side is called the *posterior*, and the first term in the numerator in the right-hand side is called the *likelihood*. The strength of Bayesian methods comes from $P(\theta)$, the *prior*. The prior summarizes knowledge about the probability of a model *before* the results, D, of an experiment are observed. When experimental data are abundant, the exact specification of the prior is usually unimportant;[26,27] when data are scarce the prior constrains the space of available models. While careless a priori assumptions may constrain the observer to the wrong part of the model space, in the analysis of genetic regulatory networks we may exploit well-established knowledge about transcriptional regulation to construct appropriate priors, as we illustrate in the section on biological priors.

Armed with the prior, we next find the a posteriori expected value of the parameters

$$\langle\theta\rangle_D = \int d\theta\,\theta\, P(\theta|D) = \frac{\int d\theta\,\theta\,P(\theta)P(D|\theta)}{P(D)}$$
$$\equiv \frac{\langle\theta\, P(D|\theta)\rangle}{\langle P(D|\theta)\rangle}\,, \qquad (6)$$

where $\langle\ldots\rangle$ and $\langle\ldots\rangle_D$ denote expectations over the prior and the posterior respectively. When the number of data N is large we expect the posterior to be tightly peaked around some value $\widehat{\theta}$ which maximizes the posterior (maximum a posteriori probability, or MAP, values), and then $\widehat{\theta}$ is the first-order term in the saddle-point asymptotic expansion of $\langle\theta\rangle_D$ in powers of $1/N$.

Even for severely undersampled problems, N is usually large enough so that $\widehat{\theta} \approx \langle\theta\rangle_D$, and it is tempting to replace

integrals in eq (6) by their saddle-point values. One of the greatest realizations in Bayesian theory in the last decades was that such replacement is wrong.[23,26,28–30] Indeed, when averaging over all possible models, each of the integrals in eq (6) will have a contribution from fluctuations around the value $\widehat{\theta}$. For example, under some very general conditions, and with an assumption that $\log P(D|\theta)$ scales linearly in the number of data, the total probability of the data, $P(D)$, has the following expansion in powers of $1/N$

$$\log P(D) = \log P(D|\widehat{\theta}) - \frac{K}{2}\log N$$
$$- \frac{1}{2}\log \det\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\bigg|_{\widehat{\theta}}\frac{\log P(D|\theta)}{N}\right] \qquad (7)$$
$$+ \log P(\widehat{\theta}) + o(N^0)\,,$$

where K is the number of parameters in the model (dimensionality of θ). The integral in the numerator of eq (6) can be written in a similar fashion. We see that the terms beyond the maximum likelihood contribution $\log P(D|\widehat{\theta})$ are generally negative and their magnitude grows with K. Thus, these terms provide a built-in punishment for model complexity. For this reason, they are known in the literature as the Occam razor.[30]

To illustrate, imagine that the prior admits two model families Θ^1 and Θ^2 with different parameters θ^1 and θ^2, such that $K_1 \equiv \dim\theta^1 < K_2 \equiv \dim\theta^2$. As an example, consider fitting a function with a polynomial of low degree (K_1) or high degree (K_2). Usually we would expect the model family with more parameters to be better at explaining the data: $P(D|\widehat{\theta^2}) > P(D|\widehat{\theta^1})$. Thus, if we were to choose a model family that explains the data best based on the maximum likelihood alone, a more complex model would win. However, the estimates within the larger model family, Θ^2, are less robust to small fluctuations, and this is picked up by the integration over all parameters; even though $P(D|\widehat{\theta^1})$ may be

smaller than $P(D|\widehat{\theta^2})$, the relation between the probabilities of the model families $P(\Theta^1|D)$ and $P(\Theta^2|D)$, as determined by eqs (6) and (7), may be different. In particular, for $N \gg 1$ the likelihood term, which scales linearly with N, always wins, and Bayesian model selection approaches that of the maximum likelihood. However, for small N the difference between the likelihood and the other terms in eq (7) is less profound, and a simpler model family, which is not the best in explaining the data, may turn out victorious.

In short, Bayes rule shows how a simpler model may be less likely, yet more probable.

Before ending our quick review of Bayesian statistics, two more comments are in order. First, as the model selection arguments are mostly important for small N, where $\log N \sim 1$, it would be a mistake to ignore $O(1)$ terms in eq (7) and use just $K/2 \log N$ as a model complexity punishment (Schwartz's Bayesian Information Criterion,[23] mentioned earlier). In particular, we believe that such replacement may be a cause of a common observation that the Bayesian Criterion overpunishes complex models.[31] Secondly, even though in this paper we mostly deal with finite parameter models, application of non-parametric methods to biological data certainly holds promise. Occam-type arguments for such cases have been discussed, for example, by Bialek and co-workers.[27,32]

Bayesian Inference of Regulatory Dynamics

A SIMPLE MODEL

Let $\mathbf{g}_t$ denote the vector of expressions (mRNA concentrations) of genes in a microarray experiment at time $t = 1, 2, \ldots, T, T+1$. The number of genes, $K_{\mathbf{g}} \equiv \dim \mathbf{g}_t$, can be of the order of thousands, but T is less than a hundred. In principle, $\mathbf{g}_{t+1}$ can depend on concentrations at all times that preceded $t+1$. However, if we view the gene expression mechanism in cells as a realization of some chemical kinetics, governed by first-order differential equations, then it is reasonable to expect that the concentrations depend only on their immediate past, i.e., $\mathbf{g}_{t+1} = \mathbf{f}(\mathbf{g}_t)$.[†] We begin with the simplest possible dynamic, a simplified version of eqs (1) and (2)

$$\mathbf{g}_{t+1} - \mathbf{g}_t = \mathbb{M}\mathbf{g}_t + \boldsymbol{\xi}_t , \qquad (8)$$
$$\langle \xi_{i,t} \xi_{j,t'} \rangle = \sigma^2 \delta_{ij} \delta_{t,t'} , \qquad (9)$$

where the noise is Gaussian, and $\mathbb{M}$ and σ are unknown. This is equivalent to

$$P(\mathbf{g}_{t+1}|\mathbf{g}_t, \mathbb{M}, \sigma) = \frac{1}{\left(2\pi\sigma^2\right)^{K_{\mathbf{g}}/2}} \exp\left[-\frac{1}{2\sigma^2}|\delta\mathbf{g}_t|^2\right] , \quad (10)$$

and

$$P(\{\mathbf{g}_{t\neq 1}\}|\mathbf{g}_1, \mathbb{M}, \sigma) = \frac{1}{\left(2\pi\sigma^2\right)^{T K_{\mathbf{g}}/2}}$$
$$\exp\left[-\frac{T}{2\sigma^2}\langle|\delta\mathbf{g}_t|^2\rangle_t\right] \qquad (11)$$

where $\delta\mathbf{g}_t \equiv \mathbf{g}_{t+1} - \mathbf{g}_t - \mathbb{M}\mathbf{g}_t$, and $\langle f_t \rangle_t \equiv 1/T \sum_{t=1}^{T} f(t)$ indicates empirical averaging over time.

Notice that unlike in eqs (1) and (2) the noise in this model is not correlated among the genes; moreover, variances are equal for all genes. Below we formulate an extension that incorporates hidden degrees of freedom (e.g., biochemistry) whose omission may otherwise lead to large or correlated noise (cf., the section on hidden control). However, with only a handful of experiments available we cannot hope that data will be able to determine millions of elements of a full covariance matrix. As data become more plentiful, it may even make sense to bypass eqs (8) and (9) completely and pursue model independent feature extraction (as formulated in the first two subsections of the Outlook section). Below we pursue the possibility that the current simple model exhibits some of the success evidenced in the Results section.

BIOLOGICAL PRIORS

Even in the simplistic form of eqs (8) and (9), the dynamic still contains too many parameters ($\sim K_{\mathbf{g}}^2$) to be tractable. We therefore search for biologically motivated priors to constrain the possible values of $\mathbb{M}$. An example of such a prior would be, for example, that genes with similar regulatory sequences should be regulated similarly. More directly, genes whose promoters have similar numbers of certain important motifs should be co-expressed, an ansatz used with notable success by Bussemaker et al.[34] in discovering regulatory regions.

This may be expressed in the following prior

$$P(\mathbb{M}|\mu, \ell) \propto \exp$$
$$\left[-\frac{\ell^2}{2} \sum_{i \neq j, k}^{K_{\mathbf{g}}} \left(\frac{M_{ik} - M_{jk}}{d_{ij}}\right)^2 - \frac{\mu^2}{2} \sum_{ij}^{K_{\mathbf{g}}} M_{ij}^2\right] , \qquad (12)$$

where the first term punishes for different regulation of genes with similar regulating sequences, and the second assures proper normalization of priors by effectively constraining the range of possible M_{ij}.[†] Here d_{ij} is a distance function measuring deviation between regulatory regions of genes i and j in terms of the number of each regulatory motif appearing, weighted by the relative importance of that motif, found by considering the entire time series as in Bussemaker et al.[34] or "quality factors" as in Bussemaker et al.[35].

The parameter ℓ plays the role of a smoothing length, as in Bialek et al.,[32] and, lacking a first-principles estimate of its value, we must integrate over ℓ, weighted by an appropriate prior. As explained in Nemenman and Bialek,[27] it is not only likely that such integration will choose the proper value of ℓ almost independently of such prior, but it may even balance a slightly improper choice of the distance measure d_{ij} and the difference form $(M_{ik} - M_{jk})^2$. The same comments relate to the mass μ and the noise variance σ as well.

[†] Support for this choice is strong, as such models have been used with great success in the design of small genetic networks (see, for example, Hasty et al.[33] for a review). Furthermore, in a study clustering genes by their dynamics, Ramoni et al.[22] tested higher-order models, but found that the first-order dynamics gave the most probable result.

[†] Physicists will recognize these as the "kinetic" and "mass" terms in a Lagrangian, respectively.

An enjoyable feature of this prior is that, like the likelihood, eq (10), it is exponentially quadratic in the unknowns $\mathbb{M}$. Thus, the posterior expectations are Gaussian integrals, which may be performed analytically using the standard Wigner current technique.[36] Following eq (6) we find for the a posteriori values of the connection matrix:

$$\langle M_{ij}\rangle_D \;=\; \left.\frac{\partial}{\partial J_{ij}}\right|_{\mathbb{J}=0} \log Z(\mathbb{J}) \;, \qquad (13)$$

$$Z(\mathbb{J}) \;=\; \int d\sigma\, d\mu\, d\ell\, P(\sigma)P(\mu)P(\ell)\left[\frac{\det \mathcal{A}}{\det(\mathcal{A}+\mathcal{G})}\right]^{1/2}$$
$$\times \exp\left[\frac{1}{2}\,(\mathbb{B}+\mathbb{J})\,(\mathcal{A}+\mathcal{G})^{-1}\,(\mathbb{B}+\mathbb{J})\right]\;. \quad (14)$$

Here the curvature tensor $\mathcal{A}$ at the saddle point of eq (12) is given by

$$A_{ij,kl} \;=\; \left[\mu^2 + \ell^2 \sum_m^{K_{\mathbf{g}}}\left(c_{mj}+c_{mi}\right)\right]\delta_{ik}\delta_{jl}$$
$$- 2\ell^2 c_{ik}\delta_{jl} \qquad (15)$$

and the time-lagged correlation, equal-time correlation, and "closeness" (the inverse of the distance matrix) matrices are

$$\mathbb{B} \;=\; \frac{1}{\sigma^2}\left\langle(\mathbf{g}_{t+1}-\mathbf{g}_t)\mathbf{g}_t^T\right\rangle_t \qquad (16)$$

$$\mathbb{G} \;=\; \frac{1}{\sigma^2}\left\langle\mathbf{g}_t\mathbf{g}_t^T\right\rangle_t\;, \quad \mathcal{G}=\mathbb{I}\otimes\mathbb{G}\;, \qquad (17)$$

$$c_{ij} \;=\; \begin{cases} d_{ij}^{-2} & \text{if } i\neq j, \\ 0 & \text{if } i=j. \end{cases} \qquad (18)$$

The mass μ regulates the integrals and is expected to be small. Then $\det \mathcal{A}/\det(\mathcal{A}+\mathcal{G})$ scales as a large positive power of $\ell^2/(\ell^2+\text{const})$, and therefore decreases as ℓ decreases. On the other hand, the exponent in eq (14) involves $(\mathcal{A}+\mathcal{G})^{-1}$, which scales as $1/(\ell^2+\text{const})$ for large ℓ. The exponent thus decreases as ℓ increases. We may then reasonably expect that the integrand in eq (14) will be peaked at some non-trivial value of ℓ; this peak should be sharp since both $\mathbb{B}$ and $\mathcal{G}$ involve a large number of samples. This may be viewed as the smoothing length selection by the data.[27]

Note that the priors over the hyperparameters σ, μ, ℓ in eq (14) are as yet undefined. Since, as mentioned above (see also Clarke and Barron[26]), their actual forms are of little importance, we may hope that by choosing them appropriately we may be able to render the integrals in eq (14) analytically tractable.

HIDDEN CONTROL

The formalism deserves experimental testing. However, one further extension offers a substantial improvement. DNA microarrays offer only a partial view into the workings of a cell, since numerous important degrees of freedom remain unobserved. In mathematical modeling of the yeast cell cycle, for example, considerable effort—see, for example, Tyson et al.[37] for a review—has been exerted to fine-tune models in which only chemistry controls the processes, and the genetic expression is a mere passive function of this chemical control. While this may or may not prove to be an accurate characterization, some chemical control unobserved by gene chip experiments certainly exists. Inclusion will clearly necessitate a model of some structure other than that of eq (8). Unlike the assumption of linearity, which merely leads to misestimation of the values of interactions, this effect may make the dynamical system appear to be not of first order, or introduce a gene-dependent noise correlation matrix and artificial couplings between genes that dominate the real ones. To avoid this, we need to supplement the vector of genes $\mathbf{g}$ with a vector of unknown hidden degrees of freedom $\mathbf{h}$. Then the evolution will take form

$$\begin{pmatrix}\mathbf{g}_{t+1}-\mathbf{g}_t \\ \mathbf{h}_{t+1}-\mathbf{h}_t\end{pmatrix} \;=\; \begin{pmatrix}\mathbb{M}^{gg} & \mathbb{M}^{gh} \\ \mathbb{M}^{hg} & \mathbb{M}^{hh}\end{pmatrix}\begin{pmatrix}\mathbf{g}_t \\ \mathbf{h}_t\end{pmatrix}+\begin{pmatrix}\boldsymbol{\xi}_t \\ \eta_t\end{pmatrix}\;, (19)$$

$$\langle\xi_{i,t}\xi_{j,t'}\rangle \;=\; \sigma^2\,\delta_{ij}\delta_{t,t'}\;, \qquad (20)$$

$$\langle\eta_t^i\eta_{t'}^j\rangle \;=\; V_{ij}\delta_{t,t'}\;. \qquad (21)$$

Within Bayesian analysis, we may integrate over the unknown degrees of freedom and their possible couplings while remaining agnostic about the identities of $\mathbf{h}$, and similarly sum over their possible dimensionality $K_{\mathbf{h}}$. As data are scarce, it is reasonable to expect that models with a small number of hidden units will dominate the posterior. Thus, we allow a full correlation matrix for the Gaussian noise in the hidden units, η_t, since this adds only a few additional parameters to our model ($\mathbb{M}^{gg}$, of course, has a few million), but allows necessary flexibility.

With such phenomenological model, the analogs of eqs (10) and (11) are

$$P(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbb{M}, \sigma, \mathbb{V}) =$$
$$\frac{1}{\left(2\pi^K \det \Sigma\right)^{1/2}}\exp\left[-\frac{\delta\mathbf{x}_t^T\,\Sigma^{-1}\,\delta\mathbf{x}_t}{2}\right]\;, \quad (22)$$

and

$$P(\{\mathbf{x}_{t\neq 1}\}|\mathbf{x}_1, \mathbb{M}, \sigma, \mathbb{V}) =$$
$$\frac{T}{\left(2\pi^K \det \Sigma\right)^{1/2}}\exp\left[-\frac{T}{2}\left\langle\delta\mathbf{x}_t^T\,\Sigma^{-1}\,\delta\mathbf{x}_t\right\rangle_t\right]\;. \quad (23)$$

Here, of course, $\mathbf{x}_t$ denotes the vector of joint expression of genes and hidden controls at the time t, and the dimensionality of this vector is K. Σ is the covariance matrix of the combined noise, which is diagonal in its gg part, has $\mathbb{V}$ for its hh part, and is zero elsewhere, and $\mathbb{M}$ is now the combined transition matrix.

Now we integrate out the unobservable concentrations of hidden regulatory units. Notice that we leave all of the coupling matrices untouched. Thus, with no information about the concentrations and identities of the controls we will still be able to later specify which genes interact with them:

$$P(\mathbf{g}_{t+1}|\mathbf{g}_t, \mathbb{M}, \sigma, \mathbb{V}, \mathbf{h}_1) =$$
$$\frac{1}{\left(2\pi^K \det \Sigma\right)^{1/2}}\exp\left[-\frac{\delta\mathbf{x}_t^T\,\Sigma^{-1}\,\delta\mathbf{x}_t}{2}\right]\;. \quad (24)$$

We will need a prior over newly introduced $\mathbb{M}$. Since chemistry couples to expression via the transcription factors,

and therefore via the regulatory sequences, we can write a similar prior as above, namely

$$P(\mathbb{M}^{gh}|\mu_{gh}, \ell) \propto$$

$$\exp\left[-\frac{\ell^2}{2}\sum_{i\neq j}^{K_{\mathbf{g}}}\sum_{k}^{K_{\mathbf{h}}}\left(\frac{M_{ik}^{gh}-M_{jk}^{gh}}{d_{ij}}\right)^2 - \frac{\mu_{gh}^2}{2}\sum_{i}^{K_{\mathbf{g}}}\sum_{j}^{K_{\mathbf{h}}}(M_{ij}^{gh})^2\right],$$

(25)

with a different mass, but the same kinetic term as in eq (12). In the absence of precise identification of the hidden degrees of freedom, we lack a biological prior on $\mathbb{M}^{hg}$, $\mathbb{M}^{hh}$ and must therefore choose a prior that does not spoil the analytic tractability of the resulting integrals, such as

$$P(\mathbb{M}^{hg(hh)}|\mu_{hg(hh)}) \propto \exp\left[-\frac{\mu_{hg(hh)}^2}{2}\sum_{i}^{K_{\mathbf{g}}}\sum_{j}^{K_{\mathbf{h}}}(M_{ij}^{hg(hh)})^2\right].$$

(26)

Performing calculations similar to those that led to eq (14) we then obtain

$$\left\langle M_{ij}^{**}\right\rangle_D = \left.\frac{\partial}{\partial J_{ij}^{**}}\right|_{\mathbb{J}=0}\log Z(\mathbb{J}),$$
(27)

$$Z(\mathbb{J}) = \int d\sigma\,d\mu\,d\ell\,P(\sigma)P(\mu)P(\ell)\left[\frac{\det\mathcal{A}}{\det(\mathcal{A}+\mathcal{G})}\right]^{1/2}$$
$$\times\exp\left[\frac{1}{2}(\mathbb{B}+\mathbb{J})(\mathcal{A}+\mathcal{G})^{-1}(\mathbb{B}+\mathbb{J})\right].$$ (28)

Here the curvature tensor $\mathcal{A}$ at the saddle point of eq (12) is given by

$$A_{ij,kl} = \left[\mu^2 + \ell^2\sum_{m}^{K_{\mathbf{g}}}\left(c_{mj}+c_{mi}\right)\right]\delta_{ik}\delta_{jl}$$
$$- 2\ell^2 c_{ik}\delta_{jl}$$ (29)

and the time-lagged correlation, equal-time correlation, and "closeness" (the inverse of the distance matrix) matrices are

$$\mathbb{B} = \frac{1}{\sigma^2}\left\langle(\mathbf{g}_{t+1}-\mathbf{g}_t)\mathbf{g}_t^T\right\rangle_t$$ (30)

$$\mathbb{G} = \frac{1}{\sigma^2}\left\langle\mathbf{g}_t\mathbf{g}_t^T\right\rangle_t, \quad \mathcal{G} = \mathbb{I}\otimes\mathbb{G},$$ (31)

$$c_{ij} = \begin{cases} d_{ij}^{-2} & \text{if } i\neq j, \\ 0 & \text{if } i=j. \end{cases}$$ (32)

SPARSENESS

Equation (28) may look easy, but it is not. It involves inverting a matrix $\mathcal{A}$, which (even if we do not count the hidden degrees of freedom as in eq (25)) has dimensions of $K_{\mathbf{g}}\times K_{\mathbf{g}}$, and is computationally intractable even for simple organisms such as yeast with their roughly 6000 genes. So if the graph of interactions between the genes is nearly complete, then no amount of experimental data or computational power will

be enough for a meaningful inference of the network's topology. We have to assume that the gene regulation networks are *sparse*. This agrees with our intuitions of some small number of agents promoting large cascades of metabolic reactions. Luckily, such an assumption seems also to be supported by some preliminary experimental data.[38] One way of enforcing sparseness is to specifically search for the network that explains the data but has the smallest number of connections; see, for example, Yeung et al.[39]

Outlook

The main difficulty in obtaining time series data is not biological or technological, but rather financial. Affymetrix chips, the more reliable of the two dominant technologies, are quite expensive. As estimated by Langmead et al.,[40] a 24-point time series with replication factor of 3 currently costs $\sim\$57,600$. As the cost of the technology decreases, and as data become more plentiful, the role of priors in inferring connectivity and possible causation diminishes.

Moreover, as noted before, with increasing data comes the possibility of model-independent, non-parametric feature extraction by learning the joint probability distributions of expression levels at different times (cf., Bialek et al.[32]). We highlight two promising such directions below.

Mutual Information and Entropy Distance

A completely model-independent visualization tool of informatics is a low-dimensional embedding of the connectivity diagram of the multiple genes via some meaningful metric. To this end, armed with a successfully learned joint probability, we may use information theory to define distance in a principled way.

An example of such a diagram based on information theoretic ideas is presented for simulated chemical kinetics in Samoilov et al.,[21] in which the mutual information

$$I(i\rightarrow j) = \int dg_i\,dg_j^+\,P(g_i, g_j^+)\log_2$$
$$\left(\frac{P(g_i, g_j^+)}{P(g_i)P(g_j)}\right)$$ (33)

was used for embedding, with the time-lagged joint probability of degrees of freedom $P(g_i, g_j^+)\equiv\langle P(g_i(t), g_j(t+\tau))\rangle_t$ learned by histogramming. While mutual information is a useful similarity measure, it is not a distance in that it does not obey the triangle inequality. However, the "entropy distance"

$$D_H(i\rightarrow j) = -\int dg_i\,dg_j^+\,P(g_i, g_j^+)\log_2$$
$$\left(\frac{P(g_i, g_j^+)^2\sigma_i\sigma_j}{P(g_i)P(g_j)}\right)$$ (34)

does obey the triangle inequality[41] and thus can be used as a metric, based on which we can form an embedding in lower dimensions and construct a process diagram as did Arkin and co-workers.[19−21] Here, by $\sigma_{\{i,j\}}$ we mean the uncertainties in measurements of $g_{\{i,j\}}$, respectively. Note that this distance is reparametrization invariant under monotonic reparametrizations $x\rightarrow f(x)$; $y\rightarrow g(y)$, $\sigma_x\rightarrow\sigma_f/f'$, etc.

Clustering by Meaningful Information: The Information Bottleneck

Clustering without identifying a specific property of interest is meaningless. In most cases, if we try to learn from data (fit a curve, select a model, extrapolate, cluster, etc.) we are doing so not to find the parameters per se, but to use them to make predictions of future data.[42] Thus, in dynamics the variables of relevance are the future gene expressions, and we should cluster data by maximally compressing them while retaining the most information about the subsequent time steps. (In contrast, genes are usually clustered by similarity of their expression levels measured in some ad hoc metric.[12,31])

This idea was put on firm information–theoretic ground recently with the development of the *information bottleneck*,[43] which, given a joint probability distribution between degrees of freedom (e.g., gene expressions at some time) and a quantity of interest (e.g., gene expressions at subsequent times), allows an iterative calculation of the meaningful clusters in a probabilistic clustering algorithm.

An important aspect omitted from current formulations of the information bottleneck is Bayesian integration over possible joint probability distributions; this procedure smoothes the data and avoids clustering noise. We expect that this will be one of the most promising as well as principled lines of research in bioinformatics.

Prognosis

The future is promising for such data-driven techniques; data are becoming more plentiful, computational power continues to exponentiate, and the data themselves are becoming more reliable, as those who hope to interpret them study more carefully their statistics and analyses. (See, for example, Naef et al.[44,45] for one such careful analysis of Affymetrix data and Affymetrix's standard data analysis.) However, before any new techniques in computational biology will be widely exploited, they must be "verified" by comparing with results agreed upon in the biological community. In this case, verification will entail corroboration of inferred causal relations among genes (or within an inferred module of genes) with the biological literature. We anticipate that such time series based techniques will find common usage as tests for consensus with known connectivities becoming more standardized, and we look forward to their continued development and implementation.

References

1. Stormo, G.D. and Tan, K., "Mining Genome Databases to Identify and Understand New Gene Regulatory Systems," Current Opinion in Microbiology, **5**, 149–153 (2002).

2. Dayan, P. and Abbott, L.F., Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems, MIT Press, Cambridge, MA (2001).

3. Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W., Spikes: Exploring the Neural Code, MIT Press, Cambridge, MA (1996).

4. Walker, M.G., Volkmuth, W., Sprinzak, E., Hodgson, D., and Kliner, T., "Prediction of Gene Function by Genome-scale Expression Analysis: Prostate Cancer-associated Genes," Genome Research, **9**, 1198–1203 (1999).

5. Golub, T.R. et al., "Molecular Classification of Cancer: Class Discover and Class Prediction by Gene Expression Monitoring," Science, **286**, 628–629 (1999).

6. Alon, U., "Broad Pattern of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," PNAS USA, **96**, 6745–6750 (1999).

7. Perou, C.M. et al., "Distinctive Gene Expression Patterns in Human Mammary Epithelial Cells and Breast Cancers," PNAS USA, **96**, 9212–9217 (1999).

8. Ross, D.T. et al., "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," Nature Genetics, **24**, 227–235 (2000).

9. Scherf, U. et al., "A Gene Expression Database for the Molecular Pharmacology of Cancer," Nature Genetics, **24**, 236–244 (2000).

10. Pinkel, D., "Cancer Cells, Chemotherapy, and Gene Clusters," Nature Genetics, **24**, 208–209 (2000).

11. Cho, R. et al., "A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle," Mol. Cell, **2**, 65–71 (1998).

12. Spellman, P.T. et al., "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces Cerecisiae by Microarray Hybridization," Mol. Biol. Cell, **9**, 3273–3297 (1998).

13. Fodor, S.P.A., Read, J.L., Pirrund, M.C., Styer, L., Lu, A.T., and Solas, D., "Light-directed, Spatially Addressable Parallel Chemical Synthesis," Science, **251**, 767–773 (1991).

14. Fodor, S.P.A., Rava, R., Huang, X.H.C., Pease, A.C., Holmes, C.P., and Adams, C.L., "Multiplexed Biochemical Assays with Biological Chips," Nature, **364**, 555–556 (1993).

15. Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., and Lockhard, D.J., "High Density Synthetic Oligonucleotide Arrays," Nature Genetics Supplement, **21**, 20–24 (1999).

16. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W., "Parallel Human Genome Analysis: Microarray-based Expression Monitoring of 1000 Genes," PNAS USA, **92**, 10,614–10,619 (1996).

17. Shalon, D., Smith, S.J., and Brown, P.O., "A DNA Microarray System for Analyzing Complex DNA Samples Using Two-color Fluorescent Probe Hybridization," Genome Research, **6**, 639–645 (1996).

18. Friedman, N., Linial, M., Nachman, I., and Pe'er, D., "Using Bayesian Networks to Analyze Expression Data," Proc. 4th Annual Int. Conf. on Computational Molecular Biology (RECOMB), 127–135 (2000).

19. Arkin, A. and Ross, J., "Statistical Construction of Chemical Reaction Mechanisms from Measured Time-series," J. Phys. Chem., **99**, 970–979 (1995).

20. Arkin, A., Shen, P., and Ross, J., "A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements," Science, **277**, 1275–1279 (1997).

21. Samoilov, M., Arkin, A., and Ross, J., "On the Deduction of Chemical Reaction Pathways from Measurements of Time Series of Concentrations," Chaos, **11**, 108–114 (2001).

22. Ramoni, M., Sebastiani, P., and Cohen, P., "Bayesian Clustering by Dynamics," Machine Learning, **47**, 91–121 (2002).

23. Schwartz, G., "Estimating the Dimension of a Model," Ann. Stat., **6**, 461–464 (1978).

24. Neumaier, A. and Schneider, T., "Estimation of Parameters and Eigenmodes of Multivariate Autoregressive Models," ACM Transactions on Mathematical Software, **27**, 27–57 (2001).

25. Press, S.J., Bayesian Statistics: Principles, Models, and Applications, Wiley, New York (1989).

26. Clarke, B.S. and Barron, A.R., "Information–theoretic Asymptotics of Bayes Methods," IEEE Trans. Inf. Theory, **36**, 453–471 (1990).

27. Nemenman, I. and Bialek, W., "Occam Factors and Model Independent Bayesian Learning of Continuous Distributions," Phys. Rev. E, **65** (2002).

28. Janes, E.T., "Inference, Method, and Decision: Towards a Bayesian Philosophy of Science," J. Am. Stat. Assoc., **74** (1979).

29. MacKay, D.J.C., "Bayesian Interpolation," Neural Comp., **4**, 415–447 (1992).

30. Balasubramanian, V., "Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions," Neural Comp., **9**, 349–368 (1997).

31. Barash, Y. and Friedman, N., "Context-specific Bayesian Clustering for Gene Expression Data," *Proc. 5th Annual Int. Conf. on Computational Molecular Biology (RECOMB), ACM Press (2001).*

32. Bialek, W., Callan, C., and Strong, S., "Field Theories for Learning Probability Distributions," *Phys. Rev. Lett.*, **77**, 4693–4697 (1996).

33. Hasty, J., McMillen, D., Isaacs, F., and Collins, J.J., "Computational Studies of Gene Regulatory Networks: In Numero Molecular Biology," *Nature Reviews Genetics*, **2**, 268–279 (2001).

34. Bussemaker, H., Siggia, E., and Li, H., "Regulatory Element Detection Using Correlation with Expression," *Nature Genetics*, **27**, 167–171 (2001).

35. Bussemaker, H.J., Li, H., and Siggia, E.D., "Building a Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis," *PNAS USA*, **97**, 10096 (2000).

36. Zinn-Justin, J., *Quantum Field Theory and Critical Phenomena*, Clarendon Press, Oxford (1996).

37. Tyson, J., Chen, C., and Novak, B., "Network Dynamics and Cell Physiology," *Nature Reviews Molecular Cell Biology*, **2**, 908–916 (2001).

38. Shen-Orr, S., Milo, R., Mangan, S., and Alan, V. "Network Motifs in the Transcriptional Regulation Network of Escherichia Coli," *Nature Genetics*, **31**, 64–68 (2002).

39. Yeung, M.K.S., Tegner, Y., and Collins, J.J., "Reverse Engineering Gene Networks Using Singular Valve Decomposition and Robust Regression," *PNAS USA*, **99**, 6163–6168 (2002).

40. Langmead, C., Yan, T., McClung, C.R., and Donald, B.R., "Phase-independent Rhythmic Analysis of Genome-wide Expression Patterns," *Proc. 6th Annual Int. Conf. on Research in Computational Molecular Biology (RECOMB), Washington DC, 18–21 April*, 205–215 (2002).

41. MacKay, D.J.C., *Information Theory, Inference and Learning Algorithms*, Cambridge University Press (2003). See http://www.inference.phy.cam.ac.uk/mackay/itprnn/.

42. Bialek, W., Nemenman, I., and Tishby, N., "Predictability, Complexity, and Learning," *Neur. Comp.*, **13**, 2409–2463 (2001).

43. Tishby, N., Pereira, F., and Bialek, W., "The Information Bottleneck Method," *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, University of Illinois Press*, 368–377 (1999).

44. Naef, F., Lim, D.A., Patil, N., and Magnasco, M.O., "DNA Hybridization to Mismatched Templates: A Chip Study," *Physical Review E*, **65**, 040902R (2002).

45. Naef, F., Lim, D.A., Patil, N., and Magnasco, M.O., "From Features to Expression: High-density Oligonucleotide Array Analysis Revisited," *Proceedings of the DIMACS Workshop on Analysis of Gene Expression Data 2001 (2002). Also e-print physics/0102010.*