Advanced Review

Informatics approaches for identifying biologic relationships in time-series data

Brett A. McKinney*

A vital goal of the genomic era is to identify biologic relationships between genes and gene products and to understand how these relationships influence phenotypes. Time course data contain a vast amount of causal and mechanistic information about complex systems, but experimental and informatics challenges must be overcome to produce and extract this information from biologic systems. Mathematical modeling and bioinformatics methods are being developed in anticipation of experiments involving the coordinated measurement of cellular and molecular quantities at various spatial and temporal scales. Experimental methods that probe at the nanoscale will facilitate the exploration of biologic systems at the single-cell and single-molecule level, but will also introduce special challenges for mathematical modeling because events at nanoscale concentrations are subject to the influence of intrinsic noise. This review addresses the progress, challenges, and frontiers in the field of time-series informatics. The ultimate goal of time-series informatics is to move beyond descriptive relationships and toward predictive models of emergent, or systemic, behaviors of biologic systems as a whole.

 Wiley Interdiscipl. Rev. Nanomed. Nanobiotechnol. 2009 1 60–68

In recent years, high-throughput, genome-wide experiments have led to the vigorous development of new bioinformatics tools and algorithms that identify genes and gene products associated with phenotypic variables and that model causal interactions in gene networks. These large-scale experiments have created vast amounts of data, but most are limited to a single snapshot in time, allowing only a coarse approximation of the underlying dynamic system. Because of the lack of time-rich biologic data, most bioinformatics efforts have focused on a static viewpoint of biology. For example, statistical methods to discriminate between phenotypic classes from microarray data are reaching maturity. However, technologies such as kinetic reverse transcription polymerase chain reaction[1,2] will soon allow for the coordinated measurement of dense time-series involving concentrations and expression levels of biologically active molecules. Bioinformatics approaches are under development in anticipation of the availability of more time-enriched data sets. This article reviews the promising developments and challenges in the area of time-series bioinformatics. These challenges include model structure inference and parameter estimation, the identification of models that predict phenotypic outcomes, and understanding the mechanisms of noise regulation in biologic systems.

Nanobased approaches are better suited than conventional methods to probe biologic systems at scales relevant to molecular mechanisms, in particular at the single-cell level.[3,4] Processes at this scale involve statistical fluctuations that may be large, and biologic systems have necessarily evolved to function in the presence of such noise. In fact, there is evidence that biologic networks exploit noise, but if not properly controlled, fluctuations may lead to inappropriate systemic behavior of the network and possibly to an increased susceptibility to disease. Modeling time series at the nanoscale, where noise effects become more pronounced, is particularly challenging but may hold important keys to understanding the etiology

*Correspondence to: Brett A. McKinney, Department of Genetics, University of Alabama School of Medicine, Birmingham, AL 35294, USA. E-mail: brett.mckinney@gmail.com

Department of Genetics, University of Alabama School of Medicine, Birmingham, AL, 35294 USA.

DOI: 10.1002/wnan.012

phenotypes that have eluded standard statistical analysis.

GRAPH-BASED NETWORKS

In this article, a model-free network is defined as a graph whose nodes represent genes and gene products and whose edges represent physical or functional interactions. The structure, or topology, of a graph provides insight into the organizing principles of cellular systems as well as into the functional motifs and interactions between genes and gene products.[5] Statistical clustering is a potentially useful way to infer graph edges by identifying the patterns of correlation between profiles in time series, but inferring network connections by correlation or other metrics may belie the underlying molecular interactions that are implicit in the time-series profiles. For example, consider the simple hypothetical time series shown in Figure 1. Profiles for quantities y and z are the most correlated and correlation-based clustering would predict a close connection between y and z, with x more distantly related. In fact, y and z co-inhibit x; however, clustering cannot capture all of the complexity of the actual model shown in Figure 2. Specifically, there is directionality and information flow to this negative feedback loop with x activating y, y activating z, z suppressing x, and each gene product degrading in proportion to its own concentration.

Clustering has limited the ability to identify the directional interactions displayed in Figure 2, but may be useful for organizing information as input for more mechanistic algorithms discussed later in this review. For clustering time series, it is important to recognize

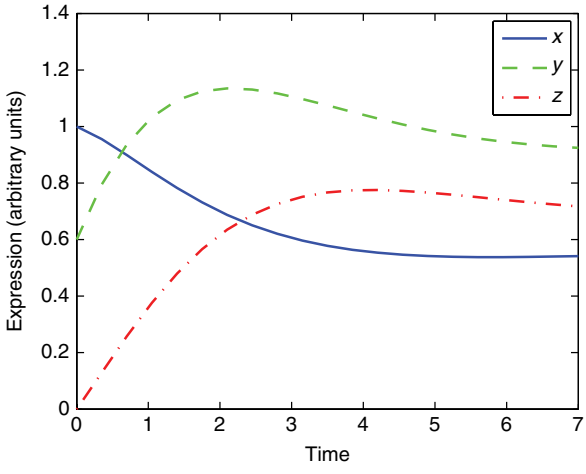

FIGURE 1 | Time-series profiles for hypothetical gene regulatory system described in Figure 2. Parameters used for the simulation are $\kappa = (0.9, 1.0, 0.6)$, $\gamma = (1.0, 0.6, 0.8)$, and $\theta = 0.9$.

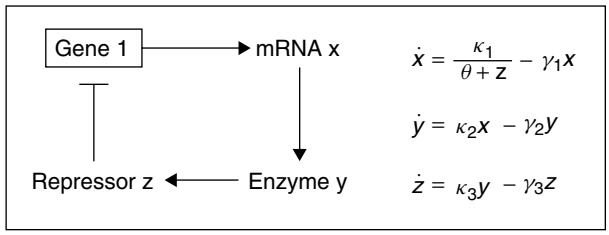

FIGURE 2 | Hypothetical single-gene regulatory network, simulated in Figure 1, involving a negative feedback loop with measured gene products x, y, and z. A single gene with mRNA concentration x produces an enzyme with concentration y. Enzyme y catalyzes a reaction step leading to metabolite z, which inhibits the gene that codes for the enzyme. Parameters κ and γ are the production and degradation constants and θ modulates the inhibitory Hill function.

that the time-series profile of each biomarker is not a collection of independent and identically distributed (iid) observations. To properly cluster time series, the dynamics should be taken into account explicitly, as for example in Ref. 6, where the authors use a linear approximation of the dynamics in the form of an autoregressive (AR) model to guide the clustering of gene expression profiles. This method should be more reliable than metric-based clustering, but AR has the disadvantage of being linear and univariate.

Graph theoretic methods facilitate the visualization of biologic relationships, but the lack of an underlying model hinders graphical methods from making experimentally verifiable predictions from system perturbations. The nodes of a realistic biologic network should represent time-varying activity and edges should represent flux through the network. At the other extreme of formalism complexity, nonlinear differential equations, which will be discussed in the following section, have been used to derive detailed models of gene networks, but the computational complexity of numerical integration and parameter estimation currently limits the size of the network that can be analyzed. The rest of the review will focus on mechanistic and data-driven modeling approaches.

MODEL-BASED NETWORKS

A Bayesian network (BN) combines probability theory and graph theory by constructing directed acyclic graphs (DAGs) that represent the dependencies between variables through probabilistic models. BNs provide a probabilistic framework for network graphs and have been used for gene network analysis of static gene expression data.[7] However, BNs are designed for time-*independent* data, and their basis in acyclic graphs prevents them from handling feedback loops, which are necessary to model real biologic networks.

In contrast to graph-based networks, which are qualitative, static representations of cellular processes and pathways, dynamic model-based networks are governed by an underlying model that allows them to generate experimentally testable output. Numerous formalisms have been used to model kinetic data, but this article will focus on more quantitative formalisms based on differential equations. For a broad overview of model formalisms, see Ref. 8.

A more reductionist approach to understanding cellular networks would be to model the noncovalent bonding and enzymatic reactions of each macromolecule. However, this would not be feasible even if precise quantitative biologic data were available; thus, in practice mathematical approximations to the physical system are used. The simplest ordinary differential equation (ODE) that may describe biologic time-series panel data is a linear system:

$$\frac{dy_i}{dt} = \sum_{j=1}^{n} A_{ij} y_j, \quad i = 1, \ldots, n \tag{1}$$

where $\boldsymbol{y}$ is a vector of functions describing the time variation of each molecule i and the constant matrix $\boldsymbol{A}$ summarizes the coupling strengths between molecules i and j. The matrix $\boldsymbol{A}$ lends itself to visualization as a graphical network, but fails to accurately model the nonlinear dynamics of a real biologic system. Equation (1) can be expressed naturally as a dynamic Bayesian network (DBN), which is a generalization of BNs to time-series data. DBN is a promising inference algorithm that has been applied to the analysis of gene networks from time series.[9–12] A DBN can also be formulated as a Kalman filter (KF),[13] a widely used tool in engineering for tracking and estimation. Using a linear system of ODEs like Eq. (1), the KF has been used to estimate DBN parameters for gene network inference.[14] The KF is a Bayesian method in the sense that it provides a way to incorporate prior information to update the current state of the system. The KF can be extended to nonlinear models by replacing the matrix in Eq. (1) with a vector of nonlinear functions as in Eq. (2) below, but the model becomes more general than a network, and the system might be better described as a dynamic Bayesian *model* or generalized DBN (GDBN). The unscented Kalman filter (UKF) is an accurate and computationally efficient method for estimating parameters of nonlinear dynamic systems.[15,16] It has been used to model *in vivo* protein time-series with noise and nonlinearities[17,18] and has the ability to model unobserved state components, yet computational limitations still must be overcome for high-dimensional systems.

Numerous classes of ODE systems have been proposed from mathematical biology to model nonlinearity in biologic systems. For example, the operon model[19,20] with nonlinear Hill functions was used to simulate the profiles shown in Figure 1. The general form of a nonlinear differential equation is

$$\frac{dy_i}{dt} = f_i[y(t), \lambda, \varepsilon_i], \quad i = 1, \ldots, n \tag{2}$$

where the model $\boldsymbol{f}$ is a vector of nonlinear functions of vector $\boldsymbol{y}$ with model parameter vector λ, and noise vector ε. Popular nonlinear ODEs for modeling biologic systems include generalized mass action (GMA) and synergistic-systems (S-systems).[21–24] The S-system preserves some of the interpretability of a purely graph-based approach while having the ability to model realistic nonlinearities. In addition, the S-system has a bounded number of parameters, making it more efficient for analyzing cellular and molecular networks than GMA. The canonical form of the S-system without noise is the following power-law system of nonlinear differential equations

$$\dot{Y}_i = \alpha_i \prod_{k=1}^{N} Y_k^{g_{ik}} - \beta_i \prod_{k=1}^{N} Y_k^{h_{ik}}, \quad i = 1, \ldots, n.$$

Each equation for the time rate of change of biochemical Y_i is composed of a term for net production from metabolic biomolecules that contribute to the increase of Y_i with rate α_i and a term representing net degradation of Y_i from catabolic biomolecules with rate β_i. The Ys may represent a molecule, cell, protein, or other gene product within the system. Kinetic order parameters g_{ik} and h_{ik}, on the real number line, represent the regulatory influence of Y_k on Y_i. In principle, the S-system reduces structure identification to parameter estimation; however, in practice, the number of parameters is too large for nonlinear systems-identification algorithms. A promising approach to reduce the computational expense of parameter estimation is to decouple the ODE system into independent algebraic equations.[25] Parameters estimated in this way may be used as initial guesses to speed up other, more computationally intensive, estimators.

It is rare that the biologic mechanism of a given process is completely described; thus, one of the goals of bioinformatics is to develop data-driven algorithms to automate the identification of the model structure and parameters from time series. The enormity of the search space of possible model structures calls for heuristic search methods such as evolutionary algorithms.[17,18,26,27] When learning the structure of

a model, it is often necessary to include parsimony constraints in the objective function. A typical choice for objective function involves some variant of least squares deviation of the model prediction from the time-series panel data. It is often useful to divide the terms in the least squares sum by the corresponding data value to prevent variables with extreme values from dominating the objective function. To penalize high-connectivity models that over-fit the data, one may add a parsimony or complexity term that is usually a function of the number of parameters in the model.[28]

SAMPLING FREQUENCY

In classical model inference, the model structure is fixed and a parameter is unidentifiable if it cannot be estimated from the data, no matter how large the sampling frequency is. The least squares definition of identifiability is often used because it takes measurement error into account.[29] Identifiability is more difficult to assess for dynamic network inference from biologic time series because the model structure is often nonlinear and/or unknown. For an experimenter, a more practical quantity is the minimum sampling frequency—or the number of time points sampled for the duration of the experiment—needed to unequivocally identify a model. If the experimental system is insufficiently sampled, the system is underdetermined, meaning multiple models may fit the data. The problem is analogous to a sample size calculation to achieve a desired statistical power in a clinical trial involving multiple regression.[30] If the structure of the model is fixed, one can minimize the reduced chi-square statistic (i.e., the maximum likelihood parameter estimation) χ^2/ν, where ν is the number of degrees of freedom, and then the level of significance can be estimated in terms of the incomplete gamma function.[31] Of course, biologic model identification typically involves the identification of the model structure as well as its parameters. It is an open research question as to how to rigorously calculate the minimum sampling frequency for biologic network identification; however, the minimum number of measurements will depend on the measurement error, the variation in the profile curvatures, the number of biomarkers in the network, and the sparseness of the connectivity of the network. For a sparse network of Boolean functions with K regulatory inputs per gene, the minimum number M of sampling points needed to identify a network of N biomarkers was shown to be of order $2^K[K + \log(N)]$.[32] This value for M was derived under ideal conditions but represents a reasonable lower bound for modeling with a more complex continuous formalism such as nonlinear differential equations.

To overcome low-frequency sampling, one could use interpolation, random effects regression, or smoothing; however, these methods could be problematic for systems with high levels of noise that cause the system to deviate from smooth profiles. A recursive approach using the UKF has been successful for parameter estimation of dynamic biologic models.[17,18] Figure 3 shows the results of this parameter estimation approach for data (filled circles) simulated based on the hypothetical model shown in Figure 2, disturbed by a large measurement noise and sparsely sampled (only five time points). The recursive steps are depicted as multiple lines for each variable shown in Figure 3 at the end of each UKF pass through the time series. In the absence of prior information on the parameters, all parameters are initialized to zero, resulting in an initial system with constant solutions. The predicted parameters at the end of each pass through the time series are used as input for the next recursion step. The UKF is insensitive to the initial choice of parameters for this model and converges to the correct parameters after 11 recursive steps. Recursion can help overcome sparsely sampled systems but also leads to increased computation time. Fewer loops through the time series and improved performance may be achieved by using high-quality initial guesses[33] or more qualitative guesses based on known pathway connectivity.

For proper experimental design, simulations should be performed to determine the necessary

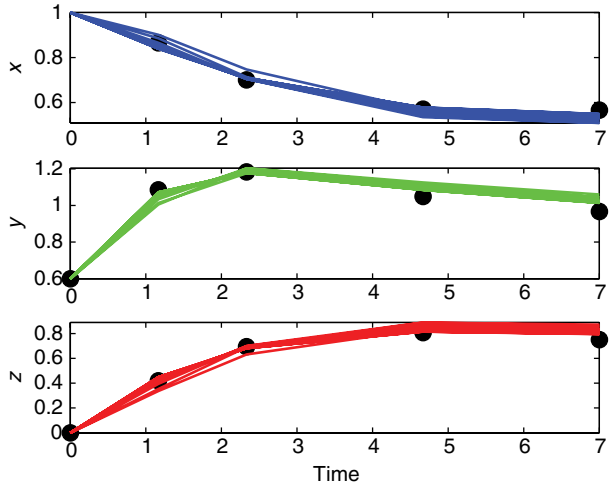

FIGURE 3 | Recursive parameter estimation with unscented Kalman filter for extremely sparse, noisy data (filled circles) simulated with model shown in Figure 2. In each panel, each overlaid predicted time curve corresponds to the recursive steps through the time series. Parameters converge after 11 steps.

sampling frequency to reduce the false-positive model rate. A true-positive detection of a dynamic model is not an all or nothing prospect; one can correctly identify parts of the model and misidentify others.[18] For stochastic system-identification algorithms, rather than taking the top-scoring system as the final model, a useful strategy to detect false-positive model components might be to inspect the *set* of top-scoring models to identify consensus model components and components that show more inter-model variation, and then run the algorithm again, this time focusing on the uncertain model components, which are more likely to be false. The best way to reduce false-positive models that all reasonably describe the available experimental data is to make computational predictions to design a new experiment that can discriminate between the hypothetical models.[34] The time-series sampling frequency need not be uniform. In a more rapidly varying domain, it is advantageous to use a higher sampling frequency in order to capture detailed features of the profile. Another practical challenge from an experimental standpoint is anticipating when such a rapid variation will occur for a given biomarker.[35] For example, transcription occurs on the scale of hours while metabolic reactions occur on the scale of minutes. Knowledge of such scales as well as time delays can aid experimental and algorithm design.

SUPERVISED MODELING

The next frontier in time-series informatics is to identify global properties and predict global states of dynamic network models. How does the perturbation of individual inputs of a noisy dynamic network affect properties of the network as a whole? Such systemic properties might be disease susceptibility or drug/vaccine response. It is not currently clear how to predict such a property directly from a dynamic network, but in other areas of bioinformatics, involving time-*independent* data, supervised statistical learning and data-mining algorithms have been used to predict the state of a phenotypic variable from multiple input variables.[36] A similar approach has been used with knowledge-driven dynamic models to simulate time-series output, which is used to train a decision tree to predict the state of a selected output variable from perturbations in initial concentrations.[37] By itself, this approach does not predict a global phenotype of an individual; however, coupled with other bioinformatics to identify gene products associated with the phenotype, the final decision leaves could predict some phenotypically relevant functional of the simulated gene product outputs. Among other things, a potential application of this technique might be the rational design of preventative and therapeutic interventions. The complexity of biologic networks poses many challenges to model-driven therapeutic design strategies due to interconnected clusters in transcription networks and the evolutionary evidence of network rewiring.[38] The robustness of gene networks to noise (Ref.39 and next section) may also make them robust to external manipulation, or may give rise to adverse side effects. Thus, a multivariate strategy is necessary to design combination therapies, which may be the best treatment strategy for many diseases.[40]

Introduced here is an integrative, supervised strategy for vaccine improvement using aspects of the dynamic model simulation method described in Ref.37. In machine learning, a problem is supervised if there is an outcome/class variable, such as a phenotype, which typically is used for classification. For rational vaccine development or improvement of existing vaccines, the goal is to maximize immunogenicity while minimizing reactogenicity. Step 1 of the proposed strategy would involve high-throughput screening to identify target cytokines associated with adverse events (e.g., Ref.41) and a parallel analysis of antibody titers to identify target cytokines associated with protective immunity. Assuming that a dynamic model exists—either knowledge or data driven—Step 2 involves the generation of a large artificial data set with random initial perturbations of cytokines and other signaling and regulatory molecules of the model as the independent variables, and the response variable is a functional involving the target-molecule (found in Step 1) expression levels at the initial and final time point. The functional acts as the outcome variable for the set of perturbations. An example functional that measures the ratio of immunogenic to reactogenic expression change from the initial to final time points for a given combination of initial perturbations p is of the form

$$F_p = \frac{\sum_{i\in\text{immunogenic}} [y_i^p(t_{\text{final}}) - y_i^p(t_{\text{initial}})]}{\sum_{r\in\text{reactogenic}} [y_r^p(t_{\text{final}}) - y_r^p(t_{\text{initial}})]} \qquad (3)$$

The next step is to find a combination of molecular perturbations that maximizes Eq. (3). Hence, Step 3 uses clustering or other methods to discretize F_p across the random Step 2 simulations into 'high', 'medium', and 'low' states, and then a decision tree is trained with the goal of identifying multivariate tree paths from the root node to 'high' output leaves. This strategy could generate hypotheses for improving protective immunity while reducing adverse events, and identify

network perturbations that lead to unstable behavior of the system. The success of this strategy is contingent upon an accurate model of the immune regulatory system and the availability of time-series data to tune such a model. An additional challenge to realizing the rational design of therapeutics is the possibly naive assumption that vaccine immune response kinetics can be modeled by the same model structure or kinetic parameters for all individuals; that is, the models may show genetic heterogeneity. A related goal will be to identify dynamic network motifs or modules[42] associated with a given phenotype and to target these motifs rationally to achieve the desired outcome.

PHENOTYPIC EFFECT OF NOISE AT NANOSCALE

When using the differential equation formalism to predict network outputs from perturbed inputs, it is commonly assumed that the concentration of each molecule or expression of each gene product varies smoothly. In reality, the expression of each molecule depends on the number and state of other molecules, which are subject to random fluctuations. These external fluctuations, or extrinsic noise, cause the number of molecules to change abruptly from time point to time point. Furthermore, isogenic, identically prepared populations of cells may vary in expression level for a particular gene due to the order of the cascade of microscopic events leading to that gene's expression. The conditions for the dominant effect of intrinsic noise in gene regulatory networks have been created in multicolor fluorescence experiments[43,44] and aspects have been modeled with detailed simulations.[45–47]

For the measurement of a molecule across M cells, the variance of the measurements is of order σ_i^2/M, where σ_i^2 is the variance of a single measurement.[48] Thus, if the measurements are averaged over a large number of cells, then one expects low noise effects and smooth time-series profiles. At the other extreme, abrupt changes in profiles may be magnified at nanoscale concentrations, where there is low copy number or low concentration. Kalman filters using the differential equation formalism can handle limited noise effects, but when statistical fluctuations of concentrations become very large, a purely stochastic formalism may be more suitable. The Gillespie algorithm has become a popular method to directly simulate the stochastic mechanisms of a dynamic system.[49] Under typical experimental conditions, it is sufficient to model with deterministic differential equations and is preferred for larger systems due to the computational cost of stochastic simulation. However, for sparsely sampled time series it may be difficult to estimate the noise strength, making it difficult to determine whether a time-series profile is merely random.

Biologic networks have evolved to function in the presence of noise and in most cases they behave in such a way as to reduce the effect of noise; however, in certain situations noise may be magnified to create heterogeneity in cell populations or to allow cells to adapt to a fluctuating environment.[45,47,50] Thus, cells and networks may exploit noise, but it is conceivable that this flexibility, if not properly controlled, may also lead to adverse systemic behavior such as disease. A possible model for the pathogenesis of some diseases that have eluded the reductionist approach to the prediction of disease susceptibility directly from the genome may be the failure of a network motif to properly regulate the intrinsic noise. Obviously environmental factors also contribute to disease, but in certain cases a disease phenotype may be a rare, emergent property of a stochastic network caused by the network's lifetime exposure to noise. For example, a stochastic mechanism has been hypothesized for haploinsufficiency diseases in which one allele in diploid cells is insufficient to assure normal function.[51] In this model, the decrease in gene dose to one allele leads to an increased susceptibility to stochastic interruptions in gene expression. This interruption may lead to the increased probability of a drop in gene expression below a critical threshold and consequently to an increase in lifetime disease susceptibility. This effect of increased noise in haploinsufficiency may also be found in tumor suppressors.[52] Another source of noise may arise from epigenetic factors, such as DNA methylation, which can modify transcriptional activity stochastically.[53] The way gene and cellular networks deal with noise and its potential role in the etiology of disease phenotypes, particularly late onset, is far from understood and represents an opportunity and challenge for time-series bioinformatics.

CONCLUSION

Biologic time series contains considerably more causal information than gene expression or protein abundance measured at a single time point; thus, the development of high-throughput technologies to gather data with high temporal information content is eagerly anticipated. However, the identification of mathematical models for these dense data poses many practical and fundamental challenges. A challenge beyond the scope of this article is the role of spatial effects as they could be important in situations such as modeling protein activity, which depends on the

protein's location within the cell. This adds another level of complexity to the computational challenges discussed above, and tools such as partial differential equations will be needed to model networks when such spatiotemporal data become more readily available. Another challenge to reaching a systems-level understanding of organisms will be to integrate other data types—genomic, proteomic, structural, environmental, clinical, and phenotypic—into time-series modeling.

A dynamic model is just that: a phenomenological *model* of the true underlying system. However, an accurate model can reveal insight into biologic relationships and may act as an *in silico* experimental tool to generate testable hypotheses. Possibly the most ambitious time-series bioinformatics research frontier is to predict global/systemic properties, such as disease susceptibility, of a biologic system from dynamic network inputs. The reductionist approach has been very successful at identifying susceptibility genes for many phenotypes, but many common multifactorial phenotypes that have eluded this reductionist strategy may be an emergent property of the entire system, as opposed to a property that is possessed by any isolated part of the system. In future predictive dynamic network models, the phenotype may be an emergent property of the model or perhaps may be modeled as a hidden variable that describes the state of the whole system. Such a model may need to include genomic, proteomic, environmental, and epigenetic factors as well as the lifetime effect of intrinsic noise expressed in regulatory networks. Such a global model of genetic and cellular networks may also lead to improved preventative and therapeutic interventions[40] by indicating ways to modulate multiple targets and simultaneously reduce adverse side effects.

NOTES

The java software used in this paper for recursive parameter estimation of generalized dynamic Bayesian networks with the unscented Kalman filter is available from the author upon request.

This work was supported by NIH Grant: K25 AI-64625.

REFERENCES

1. Baranzini SE, Mousavi P, Rio J, Caillier SJ, Stillman A, et al. Transcription-based prediction of response to IFNbeta using supervised computational methods. *PLoS Biol* 2005, 3:e2.
2. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci U S A* 1998, 95:334–339.
3. Banerjee B, Balasubramanian S, Ananthakrishna G, Ramakrishnan TV, Shivashankar GV. Tracking operator state fluctuations in gene expression in single cells. *Biophys J* 2004, 86:3052–3059.
4. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. *Science* 2002, 297:1183–1186.
5. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004, 5:101–113.
6. Ramoni MF, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A* 2002, 99:9121–9126.
7. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *Annual Conference on Research in Computational Molecular Biology (RECOMB)*. Tokyo: ACM Press; 2000, 127–135.
8. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* 2002, 9:67–103.
9. Kim SY, Imoto S, Miyano S. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief Bioinform* 2003, 4:228–235.
10. Ong IM, Glasner JD, Page D. Modelling regulatory pathways in E. coli from time series expression profiles. *Bioinformatics* 2002, 18(suppl 1):S241–S248.
11. Smith VA, Jarvis ED, Hartemink AJ. Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics* 2002, 18(suppl 1):S216–S224.
12. Smith VA, Jarvis ED, Hartemink AJ. Influence of network topology and data collection on network inference. *Pac Symp Biocomput* 2003, 11:164–175.
13. Kalman RE. A new approach to linear filtering and prediction problems. *J Basic Eng* 1960, 82:35–45.
14. Perrin BE, Ralaivola L, Mazurie A, Bottani S, Mallet J, et al. Gene networks inference using dynamic

Bayesian networks. *Bioinformatics* 2003, 19(suppl 2):II138–II148.

15. Julier S, Uhlmann J, Durrant-Whyte HF. A new approach for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Trans Autom Control* 2000, 45:477–482.
16. Voss HU, Timmer J, Kurths J. Nonlinear dynamical system identification from uncertain and indirect measurements. *Int J Bifurcat Chaos* 2004, 14:1905–1933.
17. McKinney BA, Crowe JE, Voss HU, Crooke PS, Barney N, et al. Hybrid grammar-based approach to nonlinear dynamical system identification from biological time series. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006, 73:021912.
18. McKinney BA, Tian D. Grammatical immune system evolution for reverse engineering dynamic Bayesian models. *Cancer Informatics* 2008, 6:433–447.
19. Goodwin BC *Temporal Organization in Cells*. New York: Academic Press; 1963.
20. Tyson JJ, Othmer HG. The dynamics of feedback control circuits in biochemical pathways. *Prog Theor Biol* 1978, 5:1–62.
21. Savageau MA. Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J Theor Biol* 1969, 25:365–369.
22. Savageau MA. Parameter sensitivity as a criterion for evaluating and comparing the performance of biochemical systems. *Nature* 1971, 229:542–544.
23. Savageau MA *Biochemical Systems Analysis: a Study of Function and Design in Molecular Biology*. Reading, MA: Addison-Wesley; 1976.
24. Voit EO *Computational Analysis of Biochemical Systems: a Practical Guide for Biochemists and Molecular Biologists*. Cambridge: Cambridge University Press; 2000.
25. Voit EO, Almeida J. Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics* 2004, 20:1670–1681.
26. Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics* 2003, 19:643–650.
27. Shin A, Iba H. Construction of genetic network using evolutionary algorithm and combined fitness function. *Genome Inform* 2003, 14:94–103.
28. Crampin EJ, Schnell S, McSharry PE. Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Prog Biophys Mol Biol* 2004, 86:77–112.
29. Chavent G Identifiability of parameters in the output least square formulation. In: Walter E: ed. *Identifiability of Parametric Models*. Tarrytown, NY: Pergamon; 1987.
30. Cohen J *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 1988.
31. Bevington PR, Robinson DK *Data Reduction and Error Analysis for the Physical Sciences*. 2nd ed. New York: McGraw-Hill; 1992.
32. D'haeseleer P, Liang S, Somogui R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 2000, 16:707–726.
33. Veflingstad SR, Almeida J, Voit EO. Priming nonlinear searches for pathway identification. *Theor Biol Med Model* 2004, 1:8.
34. Kremling A, Fischer S, Gadkar K, Doyle FJ, Sauter T, et al. A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. *Genome Res* 2004, 14:1773–1785.
35. Nicholson JK, Holmes E, Lindon JC, Wilson ID. The challenges of modeling mammalian biocomplexity. *Nat Biotechnol* 2004, 22:1268–1274.
36. Hastie T, Tibshirani R, Friedman J *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: New York: Springer-Verlag; 2003.
37. Hua F, Hautaniemi S, Yokoo R, Lauffenburger DA. Integrated mechanistic and data-driven modelling for multivariate analysis of signalling pathways. *J R Soc Interface* 2006, 3:515–526.
38. Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, et al. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* 2005, 309:938–940.
39. von Dassow G, Meir E, Munro EM, Odell GM. The segment polarity network is a robust developmental module. *Nature* 2000, 406:188–192.
40. Kubinyi H. Drug research: myths, hype and reality. *Nat Rev Drug Discov* 2003, 2:665–668.
41. McKinney BA, Reif DM, Rock MT, Edwards KM, Kingsmore SF, et al. Cytokine expression patterns associated with systemic adverse events following smallpox immunization. *J Infect Dis* 2006, 194:444–453.
42. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* 1999, 402:C47–C52.
43. Isaacs FJ, Blake WJ, Collins JJ. Molecular biology. Signal processing in single cells. *Science* 2005, 307:1886–1888.
44. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB. Gene regulation at the single-cell level. *Science* 2005, 307:1962–1965.
45. Kaern M, Elston TC, Blake WJ, Collins JJ. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* 2005, 6:451–464.
46. Kepler T, Elston TC. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys J* 2001, 81:3116–3136.

47. McAdams HH, Arkin A. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A* 1997, 94:814–819.

48. Wolkenhauer O, Ullah M, Kolch W, Cho K-W. Modeling and simulation of intracellular dynamics: choosing an appropriate framework. *IEEE Trans Nanobiosci* 2004, 3:200–207.

49. Gillespie DT. A rigorous derivation of the chemical master equation. *Physica A* 1992, 188:404–425.

50. McAdams HH, Arkin A. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet* 1999, 15:65–69.

51. Cook DL, Gerber AN, Tapscott SJ. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc Natl Acad Sci U S A* 1998, 95:15641–15646.

52. Kemkemer R, Schrank S, Vogel W, Gruler H, Kaufmann D. Increased noise as an effect of haploinsufficiency of the tumor-suppressor gene neurofibromatosis type 1 in vitro. *Proc Natl Acad Sci U S A* 2002, 99:13783–13788.

53. Whitelaw NC, Whitelaw E. How lifetimes shape epigenotype within and across generations. *Hum Mol Genet* 2006, 15(Spec No 2):R131–R137.

RELATED ONLINE ARTICLES

Noise in biological circuits.

