Daniel Horner

Systems Biology Applied to Hematopoietic Stem Cells: the Importance of Modelling

Abstract— **Modelling is an integral part of the systems approach to biology. Several common modelling techniques afre discussed here, including mechanistic differential equation models, traditional network models and Bayesian networks. The application of network models to the process of hematopoietic stem cell (HSC) differentiation is considered. A Bayesian networks analysis is proposed, whereby a rough structure of the cell regulatory network may be determined. Depending on these results, extensions to an existing boolean network model of HSCs can be suggested in order to incorporate signal-transduction events into the model. The new model can accommodate sustained response to a transient stimulus, a behaviour hypothesised to exist in stem cell differentiation.**

I.INTRODUCTION

A. Systems Biology

The emergence of high-throughput assays for genomics and proteomics has heralded the introduction of a new approach to biological research, termed systems biology. Proponents of the technique envision a dramatic shift in the way research is performed, underpinned by a mechanistic systems-level analysis of biological networks. Systems biology is also enabled by a emerging computational tools which permit analysis and organization of these data, and their incorporation into sophisticated models and simulations.

B.How are Models Used in Systems Biology?

Ideker et al. describe systems biology as a model-driven approach [1] with the ultimate goal of understanding the interactions responsible for the observed properties of a system. Their general framework for for systems biology begins by using available data to define an initial model. This model attempts to characterize the system to predict system response to perturbations. The second step involves the systematic application of these perturbations and collection of data from the system. Third, the original model is reconciled with experimental data through cumulative improvements or through the proposal of alternative hypotheses. Finally, the improved model suggests new perturbation experiments to further refine the representation or to distinguish between competing model hypotheses.

Clearly, different scales of model are appropriate for each stage of the process depending on the information available about the system. This paper discusses the various kinds of model available and their application to the problem of hematopoietic stem cell expansion in culture. I begin by examining some of the available model representations, their relative strengths and weaknesses, and continue by considering some examples in the context of controlling hematopoietic stem cell expansion.

C. Hematopoietic Stem Cells

Hematopoietic Stem Cells (HSCs) offer promise in the treatment of immunological and hematological diseases. They are a primitive cell type capable of large scale self-renewal and of differentiating into a variety of lineages. However, these cells differ from embryonic stem cells in that they are difficult to culture under laboratory conditions. Furthermore, there is no definitive *in vitro* assay capable of identifying this population. Rather, HSCs are identified through *in vivo* functional assays. Relying on these assays, which take several weeks to complete, is one of the challenges associated with HSC research.

Recent research has identified a combination of cytokine factors capable of sustaining these cells in culture and in stimulating a small-scale expansion in their number. Future gains may be contingent on a mechanistic understanding of the signalling pathways that implement cellular response to cytokines. Systems biology models of Hematopoietic proliferation and differentiation have much to offer in understanding and eventually controlling this regulatory mechanism.

II.Kinds of Model

There are a number of approaches to the modelling of biological processes depending on the available data and the level of representational detail desired. Two of the more commonly applied approaches are differential equations and network models.

A. Differential Equation Models

Differential equation models permit a faithful representation of many biophysical phenomena, including diffusion, enzyme-catalysed reactions and receptor-ligand binding interactions. Each chemical species, complex, and concentration is modelled according to the relevant physical law, and the solution to the resulting system of differential equations is simulated numerically.

Such generalised differential equation models suffer from computational complexity and require accurate determination of reaction kinetic parameters, which may not be readily available. In practice, simplifications are made to accommodate missing data and to reduce model complexity.

A useful variation of this model, known as the S-System model, uses a canonical differential equation representation. The rate equation for every dependent model parameter X_i is of the form[2]:

$$\frac{dX_i}{dt}=\alpha_i\prod_{j=1}^{n}X_j^{g_{ij}}-\beta_i\prod_{j=1}^{n}X_j^{h_{ij}} \qquad (1)$$

The rate parameters α and β, as well as the kinetic orders g and h are determined empirically or approximated. S-Systems can represent almost any phenomenon that can be expressed as a differential equation but have the advantage of being easier to compute and to compare due to their uniformity. The drawback of all differential equation models is the need to experimentally determine all model parameters. In specific cases where the model is of very narrow scope or the system under study is well characterized, such models perform very well.

B. Network Models

Systems biology models frequently take the form of a network. Networks are intuitively appealing to biologists because so many cellular species are interconnected: proteins in signalling cascades, metabolic pathways, gene regulatory pathways, etc. Network models are common in the fields of computer science and engineering. The last thirty years have seen them applied sporadically to biology, but their increasing adoption is indicative of their expressive potential and of the benefits to be derived from leveraging advances in network theory from other fields. They are also convenient representations of the hierarchical organization of biological systems.[3]

In a network representation, nodes generally correspond to some system state variable and edges to interconnections between variables. The model types considered below share this characteristic. The system state at any given time is captured by the values associated with each network node. Future states can be predicted using the current state and the rules associated with each network edge.

Neural Network Model

Neural networks were originally developed as simple models of brain activity. They have since been well characterised and have gained currency in many other fields due to their ability to learn. Specifically, these networks can be trained to recognize different patterns of inputs through changes in connection weighting and to respond to them with defined output values.

The network is defined as a set of nodes $y_{i \ldots n}$ and the topology is characterized by a weighting matrix w, which determines the connections between nodes. The value of a given node y_i is computed relative to the weighted values of its parent nodes y_i using some transfer function and the weights w_{ij}. One commonly used transfer function is of the following form:

$$g_i = f(y, w) = \left[1 + \exp(-\sum w_{ij} y_i + b_i)\right]^{-1} \qquad (2)$$

This sigmoidal function is advantageous because it includes a bias term b to select the sensitivity of the response on the range (0,1). The model proposed by Vohradský [5] follows this approach, with the final node activation given by a system of equations of the form:

$$\frac{dy_i}{dt} = k_{1i} g_i - k_{2i} y_i \qquad (3)$$

Here the two constants k_{1i} and k_{2i} determine the rates of gene activation and deactivation. Neural network models of this form can converge to steady state behaviour (i.e. a point attractor) or exhibit oscillations or chaotic properties.

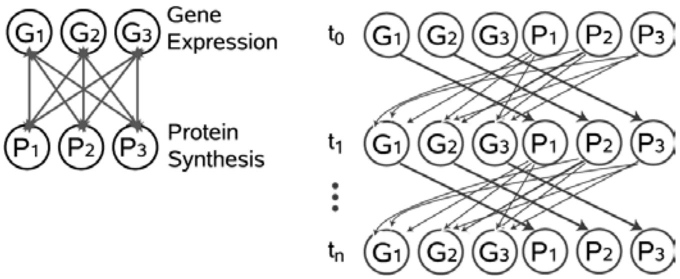

Fig. 1. Unwrapping a recurrent neural network as described in [4]. The original network models the interrelated effects of mRNA transcription (G_n) and protein translation (P_n). A linear network is obtained by expanding each node in time and only permitting forward connections.

The feedback regulation common in cellular systems results in recurrent neural network models; rather than containing a linear progression from input to output these networks are characterised by loops. Such networks are analysed by unwrapping them into a series of states at discrete time steps. This is illustrated in figure 1.

Training the network consists of determining appropriate weight matrix for the nodes given a set of data consisting of initial and final parameters for the system. A number of well-established techniques exist for achieving this, ranging from numerically solving the characteristic differential equations to back-propagation and stochastic simulated annealing methods.

The method described above requires many data points to adequately specify the system and solve the system of differential equations of the form of equation 3. Furthermore, the “unwrapping” method does not lend itself well to reactions on varying time scales, a common characteristic of biological networks; data must be available at time points of fixed period. Any interpolation will unfairly bias the weights in favour of data with longer period. A final drawback of neural network models concerns their ability to deal with missing or unobservable data points. Nevertheless, the advent of high-throughput technologies for gene and protein screening will ensure the availability of large data sets well suited to neural network analysis

Boolean Network Model

When considering genetic regulation, a common simplification restricts gene expression to Boolean values. Networks based on this simplification are a specialized instance of the neural network model described above. The model is an intuitive fit for many observed patterns of gene expression; a gene is considered to be expressed or absent at any given time with no intermediate state. This has the effect of restricting the state space of the model to to a finite domain with 2^n states.

Interactions between nodes are expressed as rules using Boolean algebra. In these networks, the discrete state-space suggests a transform of the graph to a state-machine representation where each node represents a single state (figure 2) and there is an arrow from each node uniquely identifying the subsequent state of the system. Consequently, the final state of network can be shown to relax to one of a number of terminal states termed point attractors or basin attractors. These are the discrete-level analogues of limit cycles in continuous-

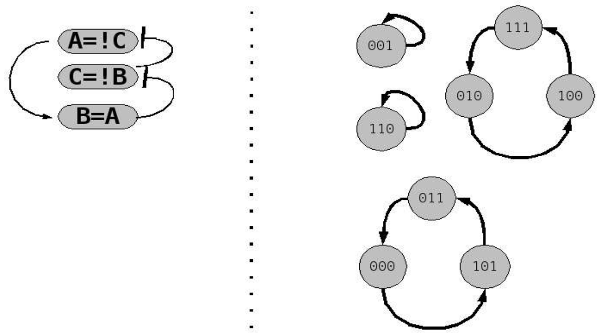

Fig. 2. Left: A simple Boolean network consisting of three genes, A,B, and C. [4] Right: The state machine transform of this network, showing two point attractors and two basin attractor.

valued networks.

Of course it is not possible to model genes which have different regulatory effects depending on their level of expression. Similar difficulties arise in attempting to model signalling pathways and protein production. Far from taking Boolean values, these phenomena may span several orders of magnitude. As a subset of the neural network model described above, Boolean networks can be trained to produce a specific output given certain input data. With adequate data this becomes a trivial operation corresponding to the discovery a “truth table” for the network. They also suffer from the same drawback as neural networks with regard to varying time scales, however the deficiency is exacerbated by the Boolean simplification: all on/off transitions must take place simultaneously. Nevertheless, the simplicity of the representation and the intriguing attractor properties make boolean networks attractive for the modelling of gene regulation networks.

Bayesian Networks

Bayesian networks differ from the network representations described above. Developed as a tool for graphical analysis and probabilistic reasoning, their origin is in statistics and operations research. However, recent work has considered their application to model discovery in biological systems [7]. In a classical Bayesian network, each node represents an observation made on the system, rather than a system property. Edges are directed from one node to another and represent causality. Put more strictly, a connection from A to B represents the conditional probability of B given A.

Like neural networks, Bayesian networks can learn to fit applied data. In the case where the network structure and all data are known, this operation finds the parameter values which maximize the likelihood of generating the training data [8]. Unlike neural networks, learning is possible even when the structure is only partially determined or when the values of some nodes are not directly observable. This has led to speculation that methods based on Bayesian networks may be suited to applications in systems biology. Additionally, the same network can accommodate Boolean and continuous-valued nodes. In the latter case, each edge is associated with a conditional probability distribution function, rather than a single conditional probability.

The “unwrapping” method described above for neural networks is equally applicable to Bayesian networks and can extend the model into the time domain. Such a network is termed a Dynamic Bayesian Network. Compared to the neural network model described above, generalised dynamic Bayesian networks offer more flexibility by permitting the time points to be chosen arbitrarily for each variable, rather than with a fixed period.

Despite the power and flexibility they offer, Bayesian networks do suffer from certain disadvantages as tools for modelling. Chief among these is a requirement for very large data sets in order to develop statistically meaningful results. Second, the generated model is probabilistic in implementation rather than mechanistic. Outcomes can be classified in terms of likelihood, but not uniquely predicted. However this may prove advantageous as many cellular processes have been shown to contain a stochastic component.

Sachs et al. have outlined a Bayesian networks approach to system structure determination [7]. By generating various candidate models for system structure, each can be fit to experimental data and assigned a Bayesian score based on the likelihood of the fit. High-scoring models are compared and features common among them are identified. Alternatively, network structure is inferred by iteratively adding or deleting edges and observing the effect on the model score. These methods have been demonstrated on both static and dynamic Bayesian network graphs using kinase activation data from the mitogen activated protein kinase signalling cascade.

Recent Systems Biology applications of Bayesian Networks have also demonstrated that data from multiple sources can be integrated, permitting cross-checking of results and expanding statistical power. In addition to these benefits, Bayesian network analysis can simultaneously accommodate dissimilar kinds of data, such as boolean and continuous-valued variables. Jansen et al. identified previously unknown protein-protein interactions by this method, pooling several data sources including mRNA co-expression data, protein function data and yeast two-hybrid interaction data. [9]

III. Modelling Hematopoietic Stem Cells

Here I investigate modelling in the context of Hematopoietic stem cells as I propose to study them. I present the system under study and examine the application of some of the network models described above to the analysis of this system

Figure 3 depicts a high-level diagram of the Hematopoietic stem-cell culture system under study. The inputs to the system $C_n(t)$ are the culture conditions in the form of cytokine growth factors. The eventual outputs are cellular response characterised as directives to proliferate (P) and/or self-renew (S). In the middle of the model are a series of measurements of cell signalling activity in the form of kinase activation within the various cellular signalling cascades (K_n).

While a proliferative response may be observed over the course of two days, true tests of stem cell self-renewal depend on *in vivo* functional assays which can take several weeks. Neither measurement is available on a time scale suitable for the feedback regulation of culture conditions. To this end, it has been proposed that measuring the activation of kinases in intracellular cascades may reveal correlations with cell pro-

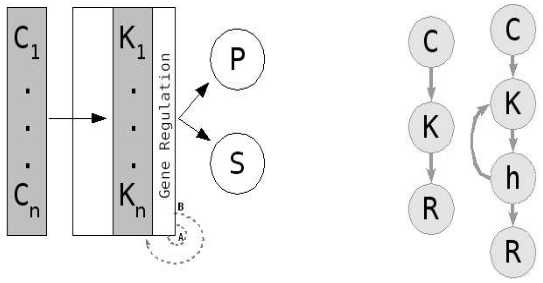

Fig. 3. In the model above, we wish to determine whether kinases are inside or outside the regulatory loop. If inside, kinase profiles will reflect input parameters. If outside, will reflect combination of input parameters and system state. Furthermore, if inside the loop there may be a kinase activation profile that can be used as a surrogate marker for the stem-cell compartment.
The bottom half of the figure depicts two possible causal relationships for the regulation of HSC fate decisions in a generalised signalling pathway. On the right, a hidden node has been added to suggest gene regulation of signalling and a dependence has been added on this node from Kinase activation. Unrolling these models into a Dynamic Bayesian Network may permit selection of the more appropriate model.

cesses.

The systems biology approach outlined above suggests we propose models for the behaviour of these cells and evaluate them by perturbation of the model parameters, i.e. the culture conditions. The model above suggests two interesting questions: First, can measurements of $K_1...K_n$ provide insight into the values of C_n conducive to a self-renewal response? If so, do $K_1...K_n$ contain any information about the decision state of the cell or alternatively, do they simply reflect a response to culture conditions? These two scenarios are represented by the dashed arrows A and B in figure 3.

A Dynamic Bayesian Network model of the process may be informative. By computing Bayesian scores for the two scenarios on the right side of figure 3, it may be possible to distinguish between these two cases.

Another model of HSC differentiation expresses gene regulation as a Boolean Network. Preissler and Kauffmann have shown [9] that even a minimal network of three simulated 'genes' can exhibit multiple basin attractors and point attractors which they associate with different compartments of differentiated cells. Their example is shown in figure 2. They suggest that a 'rule change' is responsible for the transition from one basin attractor to another, a process equivalent to lineage specification. One conceivable extension to this work replaces this abstract rule change with the specific effects of an intracellular signal. Assuming the Boolean simplification, the signal may modelled as another node in the boolean network. In the example of figure 2, a node D might be added with dependence on an external input. If a specific kinase activation can be shown to be dependent on the existing lineage commitment of the cell as discussed above, then dependencies on the state variables A-C may be added.

The original model in [9] relied on a permanent exogenous rule change to initiate differentiation. This enhancement to the model permits a sustained regulatory response to pulsed cytokine stimuli, a hypothesis of currently being tested in other stem cell lines. The new node D can be modelled to effect this rule change, perhaps even irreversibly, by adding a self-dependence or a dependence on the existing system state.

While a the simplifications inherent in a Boolean network model may not always be appropriate, it is instructive to consider them as a general framework for reasoning about cellular networks. If an observed can be reproduced in the boolean representation, it can certainly be incorporated into more sophisticated mechanistic models.

References

[1] T. Ideker, T. Galitski, and L. Hood, “A New Approach to Decoding Life: Systems Biology,” Annu. Rev. Genomics Hum. Genet. vol. 2 pp. 343-372, 2001

[2] E. O. Voit, “Computational Analysis of Biochemical Systems,” Cambridge University Press, pp. 45-65 2000.

[3] D. Bray, "Molecular Networks: The Top-Down View," Science, vol. 301, pp. 1864, 2003.

[4] H. D. Preisler, and S. Kauffman, "A proposal regarding the mechanism which underlies lineage choice during hematopoietic differentiation," Leukemia Research, vol. 23, pp. 685-694, 1999.

[5] J. Vohradsky, "Neural Model of the Genetic Network," J. Biol. Chem., vol. 276, pp. 36169-36173, 2001.

[6] J. Vohradsky, "Neural network model of gene expression," FASEB J., vol. 15, pp. 846-854, 2001.

[7] K. Sachs, D. Gifford, T. Jaakkola, P. Sorger, and D. A. Lauffenburger, (3 Sept. 2002) “Bayesian Network Approach to Cell Signaling Pathway Modeling,” Science's STKE [Online] Available: http://www.stke.org/cgi/content/full/sigtrans;2002/148/pe38

[8] K. Murphy, "A Brief Introduction to Graphical Models and Bayesian Networks," [online], 1998 [cited Oct. 26, 2003], Available: http://www.ai.mit.edu/~murphyk/Bayes/bayes.html

[9] R. Jansen, H. Yu, D. Greenbaum et al., “A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data,” Science, vol 302, pp. 449-453, 2003