

Chapter 6

Modeling the Cellular Program

The cellular program that governs the growth, development, environmental response, and evolutionary context of an organism under study does so robustly in the face of a fluctuating environment and energy sources. It integrates numerous signals about events the cell must track in order to determine which reactions to turn on, off, or slow down and speed up. These signals, which are derived both from internal processes, other cells, and changes in the extracellular medium, arrive asynchronously, and are multivalued in meaning. The cellular program also has memory of signals it has received in the past and of its own particular history as written in the complement and concentrations of chemicals contained in the cell at any instant. The circuitry that implements the working of a cell and/or collection of cells is a network of interconnected biochemical, genetic reactions, and other reaction types.

The experimental task of mapping genetic regulatory networks using genetic footprinting and two-hybrid techniques is well underway, and the kinetics of these networks is being generated at an astounding rate. Similarly, technology derivatives of genome data such as gene expression micro-arrays and *in vivo* fluorescent tagging of proteins through genetic fusion with the GFP protein can be used as a probe for network interaction and dynamics. If

the promise of the genome projects and the structural genomics effort is to be fully realized, then predictive simulation methods must be developed to make sense of this emerging experimental data. First is the problem of modeling the network structure, i.e. the nodes and connectivity defined by sets of reactions among proteins, small molecules and DNA. Second is the functional analysis of that network using simulation models built up from "functional units" describing the kinetics of the interactions. Both are necessary if the cellular program is to be understood, diagnosed when failing, and controlled.

Prediction of networks from genomic data can be approached from a number of directions. If the function of a gene can be predicted from homology, then prior knowledge of the pathways in which that function is found in various organisms can be used to predict the possible biochemical networks in which the protein participates. Similar homology approaches based on protein structural data or functional data for a protein previously characterized can be used to predict the type of kinetic behavior of a new enzyme. Thus, for example, structural prediction programs that can predict the fold of the protein product of a given gene are fundamental to the deduction of the network structure (Chapter 3 and 4).

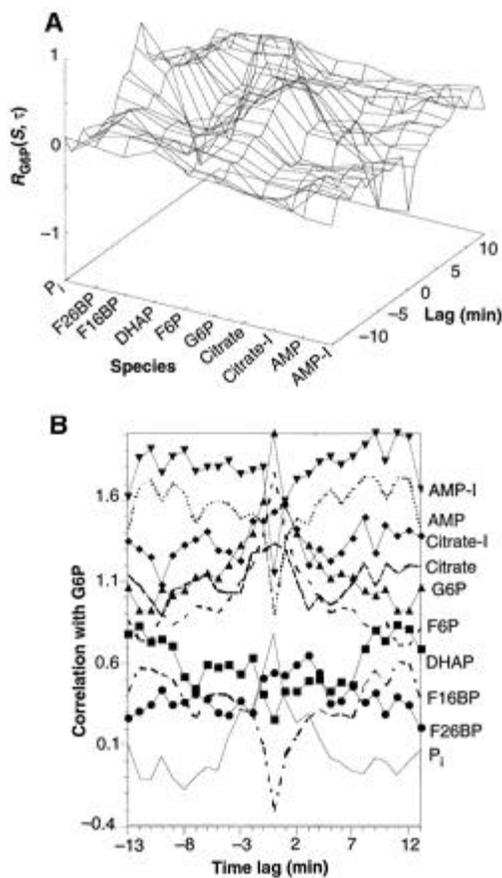
The cellular program that governs the organization of developmental pathways, metabolism, progress through the cell cycle, an organism's response to its environment, and behavior such as virulence towards another species is governed by a complex network of genetic and biochemical reactions. Recent experimental technology for obtaining time-resolved, multivariate estimates of the concentration, activities, and/or localities of cellular constituents has made it possible to observe the functioning of a biological "system" rather than simply the function of each isolated part of the cell. This opens up the possibility of developing semi-empirical models that use the experimental data to deduce network structure as well as mechanisms and kinetics. This chapter describes a transforming area of biology that, due to a rapidly expanding experimental database, will give rise to new computational challenges in the future.

Modeling the reaction pathway for the glycolytic biochemical system

A novel gene expression time series analysis algorithm known as the Correlation Metric Construction uses a time lagged correlation metric as a measure of distance between reacting species. The constructed matrix \mathbf{R} is then converted to a Euclidean distance matrix \mathbf{D} , and multidimensional scaling, MDS, is used to allow the visualization of the configuration of points in high dimensional space as a two dimensional stick and ball diagram. The goal of this algorithm is to deduce the reaction pathway underlying the response dynamics, and was used on the first few steps of the glycolytic pathway determined by experiment.

The reconstituted reaction system of the glycolytic pathway, containing eight enzymes and 14 metabolic intermediates, was kept

away from equilibrium in a continuous-flow, stirred-tank reactor. Input concentrations of adenosine monophosphate and citrate were externally varied over time, and their concentrations in the reactor and the response of eight other species were measured. The CMC algorithm showed a good prediction of the reaction pathway from the measurements in this much-studied biochemical system. Both the MDS diagram itself and the predicted reaction pathway resemble the classically determined reaction pathway. In addition, CMC measurements yield information about the underlying kinetics of the network. For example, species connected by small numbers of fast reactions were predicted to have smaller distances between them than species connected by a slow reaction.



(A) Plot of the time-lagged correlation function of G6P with all other species. The experimentally determined lagged correlation functions. The graph clarifies the temporal ordering data inherent in the correlation functions. (B) The 2D projection of the MDS diagram. Each point represents the calculated time series of a given species. The closer two points are, the higher the correlation between the respective time series. Black (gray) lines indicate negative (positive) correlation between the respective species. (C) Predicted reaction pathway derived from the CMC diagram. Its correspondence to the known mechanism is high.

Methods and Models for Deducing Genetic and Biochemical Network Structures

First we describe the problem of modeling the network connectivity using time series analysis. Most of the time series analysis techniques that have been applied to gene expression data fall into a category of statistical, distance based methods. The idea is to define a distance metric on the space of species concentrations which associates smaller distances with directly interacting species, larger distances with indirectly relating species, and very large distances with species that don't interact at all. Once a distance matrix has been constructed- an assignment of a number to each pair of species under consideration- various analysis techniques such as clustering and SNS (**define**) projection can be used to draw further meaning from the distance matrix and to represent putative interspecies relationships graphically.

The simplest distance based technique for analysing gene expression time series is that of simple correlation. The species are treated as random variables and a correlation coefficient is calculated for each pair of species and used as a measure of distance between chemical species. Simple correlation reveals linear, simultaneous relationships between variables. If two mRNA concentrations co-vary linearly, either positively or negatively, with time and/or perturbation values, this covariance will be reflected in a correlation distance measure. However, nonlinear relationships between variables are not measured by correlation coefficients, nor are time shifted linear relationships. Since gene regulation networks are thought to follow a logic best described by nonlinear hybrid algebraic differential equations, such a measure would seem to be lacking. However, the application of such a simple distance measure combined with clustering techniques have resulted in valuable and unexpected insights.

The computational cost of evaluating the correlation distance matrix with a simple

correlation distance metric is $NM^2/2$, where M is the number of genes being monitored over N time points. Since there are an estimated 100,000 genes along the human genome, calculating a distance matrix over 2000 observational time points spanning embryonic development would cost 10^{13} operations. Once a distance matrix has been constructed, analysis and visualization techniques must be applied in order to derive meaning from the distance matrix which adds additional computational overhead to the cost of the initial matrix construction (**quantify?**).

The next-simplest distance based techniques for analyzing gene expression time series use time-delayed correlations between variables at different time lags in order to construct a distance matrix. For every pair of species, a correlation coefficient is calculated for the pair at all possible time lags. In its simplest version, the distance between the two species is then taken to be the maximum correlation coefficient calculated, or some function of this maximum.

Time shifted correlations reveal linear, potentially time lagged relationships between variables. Being able to capture time shifted relationships between species is an important feature for a gene expression distance metric to have, as it allows detection of cascade-like regulation mechanisms- fairly common transcription level gene expression control structures. The simple no-lag correlation metric can miss such relationships altogether. As with the simple correlation metric described previously, nonlinear relationships between variables are not measured by time-shifted correlation coefficients. Though this is a serious limitation, time shifted correlation metrics can be considered a valuable step up in the representational hierarchy from simple correlation, as they are able to capture linear, time invariant system dynamics.

Because correlations must be calculated at

all possible time lags between variable pairs, constructing a time shifted correlation matrix is more expensive than constructing the simpler metric. If there are M genes being monitored over N time points, approximately *** (**quantify**) arithmetic operations are required to calculate the time-lagged correlation distance matrix. Calculating a distance matrix for the estimated 100,000 human genes over 100 observational time points would cost *** (**quantify**) operations. Add to that the cost of a hierarchical clustering and a total of *** (**quantify**) operations are necessary.

If all the interactions in a network were linear, then multivariate linear regression would provide the best estimate of the dependence of one variable in the system on the others. However, the dependence on the activity (or concentration) of one component as a function of the others is most often very nonlinear. In this case, linear dependency measures must be discarded in favor of general measures of dependency such as the transinformation. The transinformation is defined in terms of the joint probability distributions among sets of variables. Thus, the degree to which the value of one variable is constrained by knowledge of the values for a set of other variables is given by:

$$T(j:V) = p(j,V) \log_2 [p(j,V) / p(j)p(V)]$$

with the sum over all values of $j \times V$. There are a number of analyses that exploit this measure to produce and test network hypotheses against multivariate, often time-resolved, data.

In order to estimate the dependence of one variate on another we must calculate conditional probabilities, that is, the probability that one variable is in a one state (concentration range) given that another variate is in another state. Enough data must be collected so that the deduced relationships among variables can be deemed statistically significant.

For these analyses this amount in data can be estimated via the χ^2 statistic. If we assume

that every chemical variable in our system can take on only Q different biologically significant states, then the data constraint states that for credible analysis the minimum number of data, d , (where each data represents the observation of *all* N variables) is governed:

$$d \geq 5Q^N$$

Thus, over 5000 observations must be made for a system of ten binary variables. Obviously, this data constraint is extremely harsh for biochemical systems in which the number of biologically significant concentrations can be relatively large and the number of variables orders of magnitude greater than 10. Therefore methods (**methods or assumptions? the latter being a superposition of networks?**) must be developed for breaking large biochemical networks into smaller sub-networks which can be probed using this method.

From the time-series data a statistical analysis must predict the most probable network of interactions between chemical species that produced the observed system dynamics. To do so, the method must effectively check every possible network of connections among the measured species. While the number of such network structures rises exponentially with the number of variables composing the system, practically, the number of possible networks is greatly reduced with constraints on the solution by inserting chemical and genetic knowledge into the analysis, and to simply assume limited dependencies within the network.

Limiting the number of variables that can directly cause variations in an observation severely reduces the model space that it is necessary to test. For each variable, j , one finds the strength of the relationship between j and all other pairs of (perhaps time-lagged) variables. If the strength of the interaction is statistically significant, then retain that pair in the dependency set for the variable j . If after testing all pairs the dependency set is empty, conclude that j does not depend on any other

variable in the system. Otherwise, conclude that all variables in the dependency set are causative factors for j .

This is an $N(N-1)/2$ step algorithm (each step is composed of calculating the transinformation for each pair of variables). Each of these steps involves a three variable by M data point evaluation of a distribution estimation algorithm. All of these operations are repeated for each of the N variables, thus the scaling law is on the order of N^3M . However, the number of data points necessary to estimate the joint distributions in the transinformation for variables with Q states is of the order Q^N . The final scaling law for the

estimation becomes approximately N^3Q^N . Actually, there is some redundancy in the distribution estimation steps that might be exploited to slight reduce this Q^N dependency.

However, the assumptions behind this algorithm, that three way transinformations are enough to predict interactions of order greater than three, can lead to errors of omission in eukaryotic systems, in particular. Given that eukaryotic systems can have many multi-protein complexes containing four or more proteins, this heuristic may have to be extended to do at least four and five way interactions, $N^4Q^N - N^5Q^N$ scaling.

Methods and Models for Cellular Network Analysis

The nonlinearity of the biochemical and genetic reactions, along with the high degree of connection (sharing of substrates, products and effectors) among these reactions, make the qualitative analysis of their behavior as a network difficult. Furthermore, the small numbers of molecules involved in biochemical reactions (typical concentrations of 100 molecules/cell) ensure that thermal fluctuations in reaction rates are expected to become significant compared to the average behavior at such low concentrations. Since genetic control generally involves only one or two copies of the relevant promoters and genes per cell, this noise is expected to be even worse for genetic reactions. The inherent randomness and discreteness of these reactions can have significant macroscopic consequences such as that common inside living cells.

Chemical systems evolve with time because of changes in their constituent molecules when those molecules collide and react. Since naturally occurring molecular collisions are *random*, the temporal evolution of any chemically reacting system is *stochastic*. Elementary kinetic theory shows that, under conditions in which reactive molecular collisions are separated by many nonreactive molecular collisions, the temporal evolution of the system's state, $\mathbf{X}(t)$,

$$\mathbf{X}(t) = [X_1(t), \dots, X_N(t)]$$

constitutes a jump Markov process. That is, $\mathbf{X}(t)$ performs a "random walk" in real time over the N-dimensional integer lattice space, hopping from one lattice point to another as successive reactions occur.

An algorithm simulating jump Markov processes has been rigorously derived from the same premises that lead to the master equation (ME). The ME defines evolution of $\mathbf{X}(t)$'s probability function $P(\mathbf{x}, t | \mathbf{x}_0, t_0)$, while the simulation generates sample trajectories or "realizations" of $\mathbf{X}(t)$. The heart of the simulation is a procedure for randomly

deciding, at any time t when the system's state $\mathbf{X}(t)$ is known, at what time $t+$ the next reaction in the system will occur and which reaction, R_μ , will be the next reaction.

Using a mathematically exact procedure for generating random values for τ and μ , the simulation moves the system forward in time from one reactive collision to the next, continually updating the chemical species population levels in accordance with the outcomes of the selected reactions. The *statistical* properties of the system behavior are estimated using statistics from multiple simulations under identical conditions.

From a modeling standpoint, the simulation has two advantages over the ME: it is straightforward to apply even to complicated coupled chemical reaction schemes, and the results of the simulation are directly comparable with experimental results obtained on real systems. The primary computational bottleneck of a simulation approach to the master equation is that it can be expensive to model behavior of systems with many reacting species over extended time intervals.

There are three bottlenecks in the numerical analysis of biochemical reaction networks; the first two pertain to using the ME approach. The first is the multiple time scales involved. Since the time between biochemical reactions decreases exponentially with the total probability of a reaction per unit time, the number of computational steps to simulate a unit of biological time increases roughly exponentially as reactions are added to the system or rate constants are increased.

The second bottleneck derives from the necessity to collect sufficient statistics from many runs of the Monte-Carlo simulation to predict the phenomenon of interest. Often, such phenomena as phase-variation of coat-proteins in pathogenic virus and bacteria, oncogenesis or DNA mutation, occur at very low frequencies. Many runs of the Monte-Carlo algorithm are necessary to properly

estimate the probabilities of these events if they are to be analyzed via a stochastic simulation.

The third bottleneck is a practical one of model building and testing: hypothesis exploration, sensitivity analyses and back-calculations, will also be computationally intensive before master equation approaches can be applied to learn about the behavior of a proposed network model.

One approach to the first bottleneck problem is to develop a mode-switching algorithm that can change to numerical methods more efficient than the ME when certain conditions are met. A promising approach is to develop a simulation algorithm that "interpolates" between the master equation simulation and the standard ordinary differential equations (ODEs) used for described deterministic chemical kinetics. By first approximating the master equation by a Fokker-Planck equation (FPE), a subsequent step allows the generalization of the FPE to determine an approximate Langevin Equation (LE). An algorithm would describe the decision by which method should the dynamics should be propagated— the Monte-Carlo ME method, the LE or the ODEs. The decision would be based on criteria such as its current concentration, the concentration of those species with which it directly interacts, and the rate of the reactions in which it participates. As a simulation progresses, the mode of propagation for each species may switch, but that switching between regimes must not introduce biased errors in the integration, that neglect of the fluctuations does not lead to ablation of certain system behaviors, and that some estimate be made of the error introduced by adding heuristic submodels. This will require multiple exploratory simulations to be performed wherein the effect of varying parameters in the heuristic models are measured.

The above discussion has focused on spatially homogeneous chemical systems

where rapid mixing prevents the formation of persistent concentration gradients. Treating spatially inhomogeneous systems is more difficult. The usual deterministic approach is to convert the ordinary differential reaction rate equations into *partial* differential equations that incorporate Fick's macroscopic diffusion law. For a stochastic treatment, one has to subdivide the system volume into approximately homogeneous spatial subvolumes, and then allow diffusive exchanges of molecules between adjacent subvolumes. Stochastic simulation becomes even more computationally expensive in the spatially inhomogeneous case because of the many "diffusive exchange reactions" that must be simulated in addition to the chemical reactions.

Some stochastic simulations on spatially inhomogeneous systems have been reported, but there are unresolved technical issues associated with the choice of the subvolume size and the form of the probability rates for diffusive molecular transfers. An accelerated stochastic simulation algorithm for homogenous systems will inevitably be useful in accelerating the inhomogeneous case, since inhomogeneous systems are diffusively interacting assemblages of homogeneous subsystems.

In order to simulate the necessary statistics for a given chemical system, many runs of the Monte-Carlo algorithm must be executed, and is therefore naturally parallel. About $4(1-p)/f_e^2 p$ samples are required to estimate the probability, p , of a binary random event with 95% confidence where f_e is the desired maximum fractional error in p . Thus, a low probability event that occurs in one cell 1% of the time, and is to be predicted within +/-0.05%, would require ~10,000 simulations. Many genetic and biochemical processes are composed of tens of genes, hundreds of proteins, complexes and small molecules. In these cases the computational load is restrictive.

High End Computing Needs for Modeling the Cellular Program

Eventually the largest system of genes and proteins that will probably have to be simulated is on the order of 100 genes and regulatory elements and 500 proteins, complexes and small molecules and maybe 10 cellular compartments or locals. This is currently well-beyond the scaling laws of current simulation algorithms and 0.1 teraflop computing of today. The issues that must be addressed in this area are the disparate time scales requiring new mode-switching algorithms, and the gathering of the necessary statistics to quantify event likelihood. While the second issue unambiguously benefits from greater teraflop machines, the former does too since algorithm switching will likely only realize an order of magnitude savings in simulation time.

In addition, various computational and experimental data are vital to restricting the network structure and analysis space, i.e. more generally integrating domain knowledge into time-series analysis is required to be computationally feasible. Sources of information include not only experimental, but computational data such as large scale sequence comparisons, phylogeny information, protein fold recognition, folding and prediction of structure, enzymology and evaluation of ligand-receptor affinities and multi-protein interactions. These areas are compute-bound in their own right, and their connection to modeling of the cellular program is a natural outgrowth of the ambition of the computational biology effort in the Strategic Simulation Plan.

Requirements for Network Deduction and Analysis over the Human Genome

<u>Problem Class</u>	<u>Sustained Capability 2000</u>
Simple correlation	10¹³ flops
Time-lagged correlation	10^{**} flops
Information theoretic techniques	10^{**} flops
ME simulation (homogeneous)	10^{**} flops
ME simulation (inhomogenous)	10^{**} flops