Journal of Biological Systems

Can Markov Chain Models Mimic Biological Regulation?

SEUNGCHAN KIM*

Cancer Genetic Branch, National Human Genome Research Institute
National Institutes of Health
50 South Dr. MSC 8000, Bethesda, MD 20892, USA
dolchan@nih.gov
http://www.nhgri.nih.gov

HUAI LI

Division of Computational Bioscience, Center for Information Technology, NIH
Bethesda, MD 20892

EDWARD R. DOUGHERTY

Department of Electrical Engineering, Texas A&M University
College Station, TX 77840
Department of Pathology, University of Texas, M. D. Anderson Cancer Center
Houston, TX 77030

NANWEI CAO

Division of Computational Bioscience, Center for Information Technology, NIH
Bethesda, MD 20892

YIDONG CHEN, MICHAEL BITTNER

Cancer Genetic Branch, National Human Genome Research Institute, NIH
Bethesda, MD 20892, USA

EDWARD B. SUH

Division of Computational Bioscience, Center for Information Technology, NIH
Bethesda, MD 20892

Received (Day Month Year)
Revised (Day Month Year)

A fundamental question in biology is whether the network of interactions that regulate gene expression can be modeled by existing mathematical techniques. Studies of the ability to predict a gene's state based on the states of other genes suggest that it may be possible to abstract sufficient information to build models of the system that retain steady-state behavioral characteristics of the real system. This study tests this possibility by: (i) constructing a finite state homogenous Markov chain model using a small

*Corresponding author: dolchan@nih.gov

set of interesting genes; (ii) estimating the model parameters based on the observed experimental data; (iii) exploring the dynamics of this small genetic regulatory network by analyzing its steady-state (long-run) behavior and comparing the resulting model behavior to the observed behavior of the original system. The data used in this study are from a survey of melanoma where predictive relationships (coefficient of determination, CoD) between 587 genes from 31 samples were examined. Ten genes with strong interactive connectivity were chosen to formulate a finite state Markov chain on the basis of their role as drivers in the acquisition of an invasive phenotype in melanoma cells. Simulations with different perturbation probabilities and different iteration times were run. Following convergence of the chain to steady-state behavior, millions of samples of the results of further transitions were collected to estimate the steady-state distribution of network. In these samples, only a limited number of states possessed significant probability of occurrence. This behavior is nicely congruent with biological behavior, as cells appear to occupy only a negligible portion of the state space available to them. The model produced both some of the exact state vectors observed in the data, and also a number of state vectors that were near neighbors of the state vectors from the original data. By combining these similar states, a good representation of the observed states in the original data could be achieved. From this study, we find that, in this limited context, Markov chain simulation emulates well the dynamic behavior of a small regulatory network.

Keywords: Gene selection; Gene regulatory network; Markov chain simulation; Steady-state analysis; Melanoma.

1991 Mathematics Subject Classification: 22E46, 53C35, 57S20

1. Introduction

The use of Markov chains to enable estimation in complex models via simulation is now a widespread statistical methodology, in particular, in the context of biological systems [1,2,3]. For modeling gene regulatory networks, the most popular model is the Boolean-network model originally introduced by Kauffman [4,5]. In this model, gene expression is quantized to binary levels (on and off) and the expression level (state) of each gene at step $t+1$ is functionally related to the expression level of some other genes at step t, using logical functions. Supposing the logical function is not changed through step t, the dynamics of a Boolean network can be represented as a first-order homogenous Markov chain whose state-transition matrix is binary. The Boolean model is a degenerate Markov model in the sense that it is not stochastic, each state being deterministically dependant upon prior states. A new class of stochastic models, called Probabilistic Boolean Networks (PBNs), has recently been introduced [6]. Like Boolean networks, PBNs involve state-independent rule-based dependencies among genes; however, unlike Boolean networks, the rules themselves are random, so that PBNs are able to cope with uncertainty, both in data and model parameter selection. The dynamics of PBNs can be studied using Markov chains, with the state-transition matrix being completely specified by the many possible Boolean functions and their selection probabilities [6,7].

In this study, we take a different approach by constructing a finite-state Markov chain whose transitions depend on state-dependant multivariate conditional probabilities between gene-expression levels, based on microarray data. Mathematical

modeling tools that allow estimation of steady-state behavior in biological systems would be useful for examining two ubiquitous forms of biological system behavior. The first is homeostasis, the ability of cells to maintain their ongoing processes within the narrow ranges compatible with survival, and the second is a switch-like functionality that allows cells to rapidly transition within limited process segments between metastable states.

All cells are faced with the dauntingly complex requirement of keeping the essential processes required for cell maintenance continuously operational. In addition, they must balance the outputs of all of these processes relative to each other as adjustments are made to account for varying levels of activity of the organism, differential availability of nutrient sources, constantly variable environmental conditions and sporadic stresses. In mature organisms, most of the cells present are dedicated to performing specialized functions, and have acquired the components and machinery necessary for this terminal state of differentiation. These cells typically modulate their repertoire of expressed genes only in the minor ways required to maintain this state and carry out their specific functions.

The most detailed understanding of regulatory mechanics for a continually maintained central process is in the area of metabolism. In this domain, the ability to achieve constancy appears to be mediated at the transcriptional level by dense local regulatory connections among those elements of the system that collaborate to carry out particular synthetic or catabolic functions, with less dense connections between these local modules to adjust the outputs of each module relative to the overall requirements of the system. Models for this form of system behavior would thus be expected to rely on densely interwoven connections between collaborating elements that produce mutually reinforcing behaviors, leading to self-stabilization of a desirable target state. Since biological systems exhibit considerable stability, it would be further expected that once the system is near a target state, the existing interactions would efficiently guide the system to the target state and once there, continuously restore it to the target state after mild perturbations.

On the other hand, a complex, self-stabilizing system would not be expected to reach a desired target state starting from an arbitrary state. This difficulty of reaching a target state from a state fairly distant from the target provides a serious challenge in terms of modeling system dynamics with Markov Chains. If a model is constructed to examine what happens with rule sets abstracted from biological observations, the approach of examining the steady-state behavior achieved after many initializations from random states is unlikely to produce behavior similar to the biological system being modeled. Complex biological end-states are only reached through a very orderly progression from one highly ordered state to another. In complex multicellular organisms, this progression from egg to mature individual constitutes the process of development. At each step, the rules of interaction that govern transcriptional regulation of a particular gene could be very different. Thus, an accurate model could not be expected to use any single simple set of rules to transition the system from the many unordered states produced by random initial-

ization to the highly ordered target state from which the rules were abstracted. To make steady-state analysis after random initialization sampling meaningful, sufficient perturbation to bring the system from a random state to a state near enough to the target state for self-guidance by end-state rules to become effective would have to be utilized.

Another common system feature can be observed at the level of functional modules within biological systems. This behavior is observed as an ability to readily switch from one relatively stable state to another in response to a simple stimulus. Such an arrangement allows rapid induction and reduction of specialized activities in response to sudden demands for functions that are not used continuously or for which the extent of demand is highly variable. This ability suggests that some simple perturbations are capable of creating a cascade of regulatory interactions that can rapidly permeate a segment of a system, invoking a different set of mutually reinforcing behaviors that will drive the module toward a new target state, and then maintain it in that state. An accurate model of a biological system that switches between relatively stable states that is subjected to perturbations to allow it to reach the stable steady state distributions favored by the transition rules derived from biological observations would be expected to have a significant probability of occupying steady states similar to each of the steady states from which the rules were derived.

As a first attempt to determine whether the kinds of biological behavior described above could be captured in a Markov Chain model, a small network based on microarray data observations of a human cancer, *melanoma*, was built and simulated by a Markov Chain. This required developing criteria to select a small set of genes from which to build a Markov chain and developing a method to construct transition rules from microarray data. We then compared the model Markov Chain's steady-state behavior to the initial observations.

2. Methods

2.1. *Data set*

The gene-expression profiles used in this study result from a study of 31 melanoma samples [8]. For that study, total messenger RNA was isolated directly from melanoma biopsies, and fluorescent cDNA from the message was prepared and hybridized to a microarray containing probes for 8,150 cDNAs (representing 6,971 unique genes). Several analytical methods were applied to the expression profiles from well-measured genes to visualize the overall expression pattern relationships among the 31 cutaneous melanoma tumor samples. The clustering results indicated that the 31 melanomas could be partitioned into a major homogeneous group of 19 melanomas and a group of 12 melanomas with more varied expression behavior, as shown in Table 1. In identifying genes that discriminate these groupings of melanomas, a statistical measure was employed to generate a gene list weighted according to the gene's impact on minimizing the volume occupied by the groups

Table 1. Case number used in this study.

Exp. No.	Case No.
1	M93_007
2	M92_047
3	M91_054
4	UACC091
5	UACC502
6	UACC1097
7	UACC1256
8	UACC903
9	UACC1273
10	UACC930
11	UACC2837
12	UACC827 T
13	WM1791C
14	UACC647
15	UACC2534
16	M92_001
17	UACC457
18	HA_A
19	UACC383
20	UACC3093
21	A_375
22	UACC1529
23	UACC1022
24	UACC1012
25	TC_F027
26	TD1348
27	TC_1376_3
28	TD_1376_3
29	TD_1730
30	TD_1638
31	TD_1720

and maximizing center-to-center inter-group distance. A study of the most highly weighted genes in this list identified a particular signaling molecule, WNT5A, and a central signal transduction pathway that appear to invoke an invasive phenotype in melanoma cells [9]. The effect of increasing or decreasing the activity of WNT5A appears to initiate a large cascade of changes in the transcription and activation states of other genes in a reversible fashion. As such shifts between metastable states seem particularly suited to Markov Chain simulation, genes from this pathway were chosen as a nucleus of the model system. Further genes for the model were chosen from a set of 587 genes from the melanoma data set that have been subjected to an analysis of their ability to cross-predict each other's state in a multivariate setting [10,11]. For the purposes of this analysis, each gene's expression level was quantized to a ternary value that represents the abundance of messenger RNA produced by that gene in a particular melanoma sample relative to the abundance of messenger RNA produced by that gene in a reference cell. The val-

ues are over-expressed, equivalently-expressed, and under-expressed, relative to the reference.

2.2. *Gene selection for Markov Chain simulation*

cDNA microarrays are capable of profiling gene expression patterns of tens of thousands of genes in a single experiment. However, it would be unrealistic to investigate all genes in one regulatory network for several reasons: (1) the size of network is so large that no mathematical or computational tools can handle such a task, and (2) in many cases there exists a regulatory sub-network in which only a small number of genes are actively interactive with each other. Therefore, we set out to choose a small set of genes for which both microarray data and some biological characterizations are available to guide finite-state Markov-chain modeling. General criteria to select important genes are: (1) their predictive relationships based on coefficient of determination (CoD) analysis [10,11], (2) their roles in classifying malignant melanoma [8], and (3) their biological functionalities.

The first set of genes was chosen based on the results of multivariate measurement of gene expression relationships [12], which find associations between the expression patterns of individual genes by determining whether knowledge of the transcriptional levels of a small set of genes can be used to predict the transcriptional state of another gene. For this study, from the microarray data, we estimated coefficient of determinations of single-, two-, and three-gene predictors for a set of 587 well-measured target genes that show sufficient changes in expression values over the set of 31 melanoma samples to be useful as state predictors. As close interconnectivity between network components is desired, genes capable of both predicting other genes well and being well predicted by other genes were chosen.

Since the demand for computational resources for CoD analysis is immense, a high-performance parallel computing facility and parallel database system were utilized. First, we identified a group of predictors that can simultaneously predict multiple target genes. The more target genes a set of predictive genes can predict well, the larger is its extent of prediction. Then, we also located genes that can be well predicted by many genes. Strong inter-predictability, prediction of as well as prediction by, members of a small set of genes is taken as an indicator that these genes are highly coupled at the regulatory level and that the regulation acting on these genes is directed at achieving the goals of a particular functional segment of the network. Taking the intersection of these two gene sets both meets this requirement and reduces the number of candidates for the network. Further requirements for this core group of genes (alone or in combination) are that they should (1) show characterized biological functionalities; (2) control and regulate the activity of other genes; (3) modulate the phenotype of a cell. Genes meeting these criteria were selected to be modeled by Markov chain simulation.

2.3. *Formulation of Markov Chain model*

The Markov chain model contains n nodes, of which each node represents one of the n genes selected. Each gene has a ternary value, which is assigned from over-expressed (1), equivalently-expressed (0), and under-expressed (-1). The state space of the Markov chain has 3^n states. For capturing the dynamics of the network, we consider a "wiring rule" such that the expression state of each gene at step $t+1$ is predicted by the expression levels of genes at step t in the same network. For each target gene, a set of three predictor genes is chosen with the highest CoD value. Instead of using many possible Boolean functions that are independent of the state of the system, as in the PBN model, we use the states of three predictor genes at step t and the corresponding conditional probabilities, which are estimated from observed data, to derive the state of target gene at step $t+1$. Eq. 2 shows the definition of the Markov chain between a state at step t and the state at step $t+1$,

$$\mathbf{S}^{(t)} =: \left(g_1^{(t)} g_2^{(t)} \dots g_n^{(t)}\right) \rightarrow \mathbf{S}^{(t+1)} =: \left(g_1^{(t+1)} g_2^{(t+1)} \dots g_n^{(t+1)}\right). \tag{1}$$

The transition rule is depicted in Figure 1 and characterized by

$$g_l^{(t+1)} = \begin{cases} -1 : \text{with } C_l^{-1}\left(g_i^{(t)} g_j^{(t)} g_k^{(t)}\right) = \Pr\left(g_l^{(t+1)} = -1 \middle| g_i^{(t)} g_j^{(t)} g_k^{(t)}\right) \\ 0 : \text{with } C_l^{0}\left(g_i^{(t)} g_j^{(t)} g_k^{(t)}\right) = \Pr\left(g_l^{(t+1)} = 0 \middle| g_i^{(t)} g_j^{(t)} g_k^{(t)}\right) \\ 1 : \text{with } C_l^{1}\left(g_i^{(t)} g_j^{(t)} g_k^{(t)}\right) = \Pr\left(g_l^{(t+1)} = 1 \middle| g_i^{(t)} g_j^{(t)} g_k^{(t)}\right) \end{cases} \tag{2}$$

where $i, j, k, l \in \{1, 2, \dots, n\}$, $C_l^{-1} + C_l^0 + C_l^1 = 0$, and C_l^{-1}, C_l^0, and C_l^1 are the conditional probabilities, which depend on the states of the predictor genes. For three predictor genes for a target gene with a ternary value, there are $3^3 = 27$ possible states observable. The conditional probabilities C_l^{-1}, C_l^0, and C_l^1 are estimated from the data. Since the number of experiments (data) in microarray studies is often limited, there may be some states not observed in the data. In such case, we assign $\Pr(g_l = -1)$, $\Pr(g_l = 0)$, and $\Pr(g_l = 1)$ for C_l^{-1}, C_l^0, and C_l^1, respectively. Based on the transition rule, we can compute the transition probability between any two arbitrary states of the Markov chain as follows:

$$\Pr\left\{\mathbf{S}^{(t)} \rightarrow \mathbf{S}^{(t+1)}\right\} = \prod_{l=1}^{n} C_l^{g_l^{(t+1)}}. \tag{3}$$

2.4. *Implementation and simulation of the Markov Chain model*

After constructing the Markov chain based on multivariate relationships among genes inferred from coefficient of determination analysis, conditional probabilities of n three-predictor sets for each possible state are estimated from the data. An important consideration is whether or not there exists a steady-state distribution for the chain. The Markov chain is said to have a steady-state distribution if there exists a probability distribution $p = (p_1, p_2, \dots, p_M)$ such that for all states

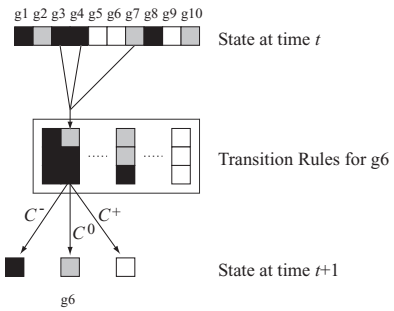


Fig. 1. The structure of the Markov chain model

$i, j \in \{1, 2, \ldots, M\}$, $\lim_{r\to\infty} P_{ij}^r = \pi_j$, where P_{ij}^r is the r-step transition probability. If there exists a steady-state distribution, then regardless of the initial state, the probability of the Markov chain being in state i in the long-run can be estimated by sampling the observed states in the simulation. As with PBN, to guarantee collecting useful information from the distribution of interest, gene perturbation is added to make the chain become ergodic [7]. For a finite-state chain, ergodicity implies it possess a steady-state distribution. For ternary gene expression, the random gene perturbations can be formulated as follows. Define a perturbation flag vector $\gamma \in \{0, 1\}^n$, $n = 10$. For simplicity, we can assume the components of γ to be independent and identically distributed (i.i.d). Thus, $\Pr\{\gamma_l = 1\} = p$ for all $l = 1, \ldots, n$, where p is the perturbation probability. Suppose that at every step of the transition, we have a realization of γ. If $\gamma_l = 1$, then the state of gene g_l needs to be changed in the following way:

- If $g_l = -1$, change to "0" with probability 0.5 or "1" with probability 0.5;
- If $g_l = 0$, change to "-1" with probability 0.5 or "1" with probability 0.5;
- If $g_l = 1$, change to "-1" with probability 0.5 or "0" with probability 0.5;

The expression values of other genes remain unchanged. Otherwise, the state transition follows the transition rules given. Considering gene perturbation, we need to generalize the computation of transition probability. Let us assume the perturbation flag vector γ is i.i.d. and the gene g_l which has q-nary expression values will change to its new value with uniform probability p_0. Then, Eq. 3 can be generalized as:

$$\begin{aligned}\Pr\{\mathbf{S}^{(t)} \to \mathbf{S}^{(t+1)}\} \\ = \left(\prod_{l=1}^{n} C_l^{g_l^{(t+1)}}\right) \times (1-p)^n + p^{n_0}\,(1-p)^{n-n_0}\,p_0^{n_0} \times \mathbf{1}_{\left[\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}\right]} \quad (4)\end{aligned}$$

where p is the perturbation probability for each gene, and n_0 is the number of genes to be perturbed as given below:

$$n_0 = \sum_{l=1}^{n} \mathbf{1}_{\left[g_l^{(t)} \neq g_l^{(t+1)}\right]},$$

and $p_0 = 1/(q-1)$. In ternary case, $q = 3$, so $p_0 = 0.5$. The two terms in Eq. 4 correspond to the two cases of the next state either given by regular transition rule or by perturbation rule. The event of $\gamma = (0, \ldots, 0)$ (i.e. no gene is perturbed) occurs with probability $(1-p)^n$, it gives the first term in Eq. 4. If at least one gene is perturbed, then the transition probability depends on the number of perturbed genes n_0. Since we assume $\gamma \in \{0,1\}^n$ is i.i.d., the probability that $\mathbf{S}^{(t)}$ be changed to $\mathbf{S}^{(t+1)}$ is equal to $p^{n_0}(1-p)^{n-n_0} p_0^{n_0} \times \mathbf{1}_{\left[\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}\right]}$ which is a second term in Eq. 4. Given any arbitrary state, we computed the transition probabilities to all possible states in the chain. The summation of those probabilities is equal to 1[a]. It verified that the matrix generated by Eq. 4 is a Markov transition matrix. The simulation algorithm used in this study can be summarized as follows:

Step 1 randomly initialize $\mathbf{S}^{(0)}$

Step 2 start to run the Markov chain from $\mathbf{S}^{(t)}$ to $\mathbf{S}^{(t+1)}$ based on the following transition rule:

```
if perturbation flag = true then
  derive S^(t+1) using perturbation rule
else
  use known conditional probabilities to derive S^(t+1)
```

Step 3 repeat Step 2 for T iterations.

Step 4 start to collect sample from $\mathbf{S}^{(T+1)}$ to $\mathbf{S}^{(T+N)}$.

Step 5 repeat Step 1 through Step 4 R times, randomly initializing $\mathbf{S}^{(0)}$.

Step 6 analyze the averaged histogram distribution of R histograms for all possible states.

In the simulation, we first use the Kolmogorov-Smirnov statistic [13,14,15] to diagnose whether the chain converges after T iterations. Given Markov chain samples $\mathbf{S}^{(T+1)}, \mathbf{S}^{(T+2)}, \ldots, \mathbf{S}^{(T+N)}$, we want to compare the distributions of the two halves of these samples, $\mathbf{S}^{(T+1)}, \mathbf{S}^{(T+2)}, \ldots, \mathbf{S}^{(T+N/2)}$ and $\mathbf{S}^{(T+N/2+1)}, \mathbf{S}^{(T+N/2+2)}, \ldots, \mathbf{S}^{(T+N)}$. Since the test is devised in terms of i.i.d. samples, there needs to be a correction for the correlation between the $\mathbf{S}^{(t)}$'s. This correction can be achieved by sub-sampling the data with the interval G. For each of the two halves above, we select M sub-samples $\mathbf{S}_1^{(G)}, \mathbf{S}_1^{(2G)}, \ldots, \mathbf{S}_1^{(MG)}$ and $\mathbf{S}_2^{(G)}, \mathbf{S}_2^{(2G)}, \ldots, \mathbf{S}_2^{(MG)}$. The Kolmogorov-Smirnov statistic is defined as the maximum value of the absolute difference between two cumulative distributions and can

[a]The proof is given in Appendix A

be described as:

$$K = \max_{0 \le x < 3^n} |F_1(x) - F_2(x)| = \frac{1}{M} \max_{0 \le x < 3^n} \left| \sum_{k=1}^{M} \mathrm{I}_x \left(\mathbf{S}_1^{(kG)} \right) - \sum_{k=1}^{M} \mathrm{I}_x \left(\mathbf{S}_2^{(kG)} \right) \right| \quad (5)$$

where $\mathrm{I}_x(\cdot)$ is the indicator function. The distribution of K in the case of the null hypothesis (two data samples are drawn from the same distribution) can be calculated, thus giving the significance level probability for the null hypothesis. For analyzing the steady-state (long-run) behavior of the chain, simulations with different perturbation probabilities are run.

On each run, the chain is restarted R times with different initial states and the global relative entropy (GRE) between the histogram distributions of two restarting times is computed to measure the closeness of the two distributions. The GRE is given by

$$GRE(p_1, p_2) = \sum_x p_1(x) \log \frac{p_1(x)}{p_2(x)} \quad (6)$$

where $p_1(x)$ and $p_2(x)$ are two histogram distributions obtained from two restarting times. After simulations, taking the average of the R histogram distributions collected provides our estimate of the steady-state distribution.

3. Simulation Results and Discussion

3.1. *Genes selected for the simulation and transition rules*

Based on the coefficients of determination between each target gene and many possible predictors, the first set of fifty genes was selected by considering them to be part of both good predictors and good targets. These fifty genes were further reduced to a set of ten genes on the basis of either their known or likely roles in the WNT5A driven induction of an invasive phenotype in melanoma cells, or their close predictive relationships with these genes, as shown in Tables 2 and 3. Note that the gene *pirin* was not in the fifty-gene set. It was only included based on its high discriminative weight in a previous analysis [8] and the fact that it was a very good predictor for many targets. For these selected genes, we estimated CoDs of single-, two-, and three-gene predictors from the data. The highest CoDs for each target are shown in Table 4. Based on Table 4, we obtain the wiring diagram shown conceptually in Figure 2. Each gene has three arcs coming in, but may have more (or less) arcs going out. The thickness of an arc as well as the distance between selected genes shows the strength of relationship, i.e., CoD. The thicker the line or the closer the genes, the stronger the relationship. For example, WNT5A and pirin have a strong relationship to each other. Between pirin and STC2 we see strong predictability from STC2 to pirin, but not in the other direction. Also, note that WNT5A and pirin have many arcs outbound while PHO-C only has only one. This diagram is the result of a tremendous amount of abstraction, and is not intended as an explicit mechanistic diagram of the relationships. The influences

Table 2. Forty one of fifty mutually interactive genes selected from 587 genes.

IMAGE Clone ID	Gene Description
208718	annexin A1
760224	X-ray repair complementing defective repair in Chinese hamster cells 1
86017	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
364975	ESTs
52489	fibroblast growth factor 9 (glia-activating factor)
754479	Hypothetical protein, expressed in osteoblast
510130	cadherin 17, LI cadherin (liver-intestine)
357278	ESTs
814615	methylene tetrahydrofolate dehydrogenase (NAD+ dependent), methenyltetrahydrofolate cyclohydrolase
512472	ferritin, light polypeptide
70692	plasminogen activator inhibitor, type II (arginine-serpin)
309864	jun B proto-oncogene
627114	Killer cell lectin-like receptor subfamily C, member 2
37796	secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1)
810282	inositol 1,3,4-triphosphate 5/6 kinase
366971	topoisomerase (DNA) II alpha (170kD)
51432	RAD23 (S. cerevisiae) homolog A
897788	protein tyrosine phosphatase, receptor type, F
813673	Human mRNA for hepatoma-derived growth factor, complete cds
208001	CD59 antigen p18-20 (antigen identified by monoclonal antibodies 16.3A5, EJ16, EJ30, EL32 and G344)
754358	ESTs
135454	ESTs
746321	ESTs
823696	interferon-induced protein 56
49950	flap structure-specific endonuclease 1
138936	erythrocyte membrane protein band 7.2 (stomatin)
322537	ESTs
293328	Homo sapiens clone 24859 mRNA sequence
68977	proteasome (prosome, macropain) subunit, beta type, 10
108837	small inducible cytokine A2 (monocyte chemotactic protein 1, homologous to mouse Sig-je)
36844	interleukin 1 receptor antagonist
471631	transcription factor 8 (represses interleukin 2 expression)
264576	ESTs
823590	Sialyltransferase
130895	ESTs
774036	glutathione-S-transferase like
44311	melanoma adhesion molecule
286249	heat shock 70kD protein 1
897806	Hypoxia-inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
549146	stimulated trans-acting factor (50 kDa)
813841	plasminogen activator, tissue

diagrammed may be the result of many intervening steps that are not shown, and influence in both directions may simply reflect such a tight coupling that no basis

Table 3. *pirin* and nine of fifty mutually interactive genes selected from 587 genes used in the simulation.

IMAGE Clone ID	Gene Card Symbol	Gene Description
234237	pirin	pirin*
324901	WNT5A	wingless-type MMTV integration site family, member 5A
759948	S100	S100 calcium-binding protein, beta (neural)
25485	RET-1	reticulon 1
324700	MMP-3	matrix metalloproteinase 3 (stromelysin 1, progelatinase)
43129	PHO-C	phospholipase C, gamma 1 (formerly subtype 148)
266361	MART-1	melan-A
108208	HADHB	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), beta subunit
40764	synuclein	synuclein, alpha (non A4 component of amyloid precursor)
130057	STC2	stanniocalcin 2

Table 4. The CoD values of the highest 3-gene predictor for 10 target genes

Predictor 1	Predictor 2	Predictor 3	*Target*	Coefficient of Determination
WNT5A	STC2	HADHB	*pirin*	0.709
pirin	S100P	RET-1	*WNT5A*	0.683
WNT5A	RET-1	Synuclein	*S100P*	0.795
pirin	WNT5A	S100P	*RET-1*	0.625
S100P	RET-1	HADHB	*MMP-3*	0.700
MART-1	synuclein	STC2	*PHO-C*	0.920
pirin	WNT5A	MMP-3	*MART-1*	0.793
pirin	WNT5A	MMP-3	*HADHB*	0.772
pirin	S100P	MART-1	*synuclein*	0.559
pirin	WNT5A	PHO-C	*STC2*	0.479

for estimating directionality is available. Some generalizations that emerge from the diagrams, such as the wide influence of the state of WNT5A on the states of other genes are expected to be true.

Based on the selected predictors for each target, we infer transitional rules between states using conditional probabilities to determine state transitions.

3.2. *Steady-state behavior*

This study focused on the steady-state behavior of the Markov chain constructed from the multivariate relationships and the transitional rules, both estimated from the data. After constructing a Markov chain, it is run on simulation with the parameters, $T = 2,000,000$, $N = 6,000,000$, and $R = 500$. Computing time for each run is about 3 hours on a 1.4GHz PC. To guarantee a steady-state distribution, different perturbation probabilities, $p = 0.001$, 0.005, 0.01, 0.05, and 0.1 are used. First, we present the Kolmogorov-Smirov (KS) test and the global relative entropy

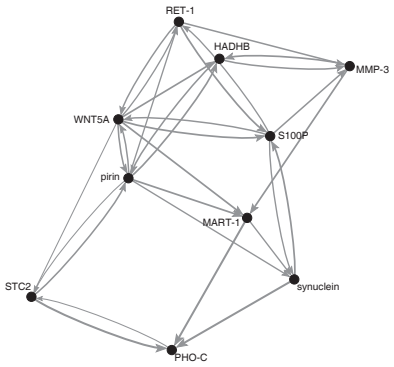
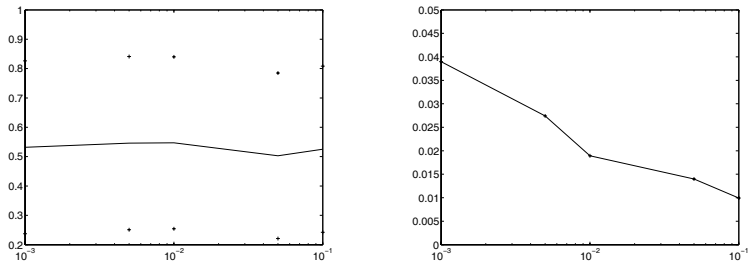


Fig. 2. Multivariate relationship between genes

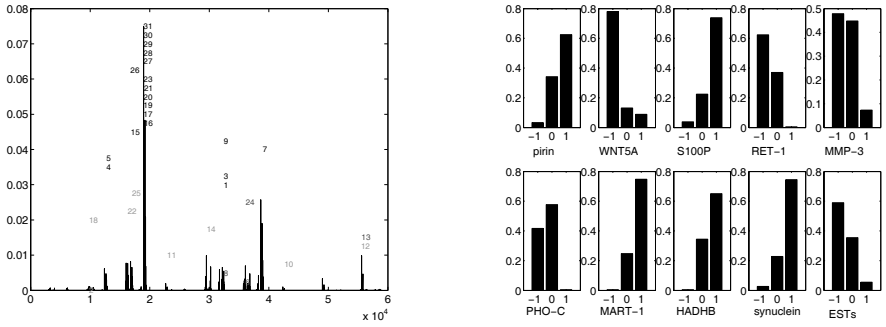
(GRE) for each perturbation probability to verify the convergence of Markov chain after T transitions. In the Kolmogorov-Smirov test, we chose the sampling interval, $G = 10$. Figure 3(a) shows the means and the standard deviations of the significance probability in the KS hypothesis test. As we can see from Fig. 3(a), all means are larger than the significance level 0.05, which implies that we can always accept the null hypothesis. Fig. 3(b) shows the means and the standard deviations of GRE. We can see that the value of means and the standard deviations of GRE are very small, which tells us that the distance between the two distributions is very close. We also observe that the value of the means of GRE decreased when the perturbation probability becomes stronger. It indicates that a more accurate steady-state distribution can be obtained with stronger perturbation. However, if perturbation is too strong, the model structure will be destroyed. We illustrate the results with stronger perturbation in Figure 5. Based on the above results, we conclude that after T transitions, the Markov chain reaches a steady-state. Hence, the distributions sampled during another N iterations estimate the steady-state distribution.

We should notice that there exist, in the steady-state distributions, only a small number of states have significant probabilities and most of those states with high probability are observed in the data. In Table 5, the ranks of each observation in terms of steady-state probability are shown, for different perturbation probabilities, p. All the observations stayed within the top 6.7%, 6.1% and 7.7% of ranks for each p. In the table, even those states with higher ranks but not observed in the data are in fact very close to the observed data. We computed the ternary-valued Hamming distance, which is simply a sum of bit-wise differences between states, and we found more than 85% of those states with high steady-state probability but not observed are within 4 Hamming distance from the observed data, which means two to four genes are at different states. While perturbation of state is necessary to guarantee the existence of a steady-state distribution, its use should be with care. Figs. 4 and 5 show steady-state distributions (part a) and the state distribution of each



(a) (b)

Fig. 3. The mean and deviation of the significance probability for KS test (a) and the global relative entropy (GRE) (b) under different perturbation probabilities. The chain was restarted 500 times with different initial state and the significance probability of KS hypothesis test and GRE was calculated each time.



(a) (b)

Fig. 4. The estimated histogram distribution after long run. (a) The steady-state distribution of all possible states of the chain with perturbation probability $p = 0.001$. (b) The marginal histogram distribution for each gene.

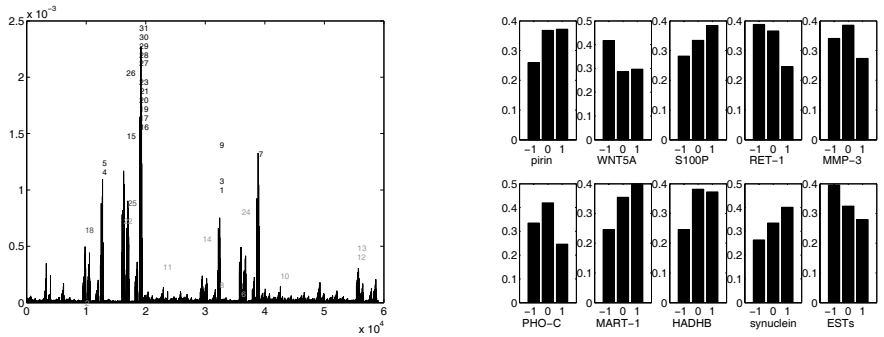
gene (part b), for $p = 0.001$ and $p = 0.1$, respectively. When the perturbation probability is small but enough to guarantee a steady-state distribution, it still keeps the structure of our observation from microarrays, but when the perturbation probability becomes too large, it destroys the structure. In Fig. 5(a), we find with significant probability many states that were not observed in the data, and the state distributions of each gene tend to flatten out in Fig. 5(b), which is significantly different from the data.

With small perturbation probability, but enough to obtain the steady-state distribution, we can appreciate the transitions of state space in the absence of perturbation. Figure 6 shows two such state-transition diagrams. In the diagram, the

Table 5. Ranks of the observed states in 31 experiments $p = 0.001$, 0.01, and 0.1

Case No.	pirin	WNT5A	S100P	RET1	MMP3	PHOC	MART1	HADHB	synuclein	STC2	Rank for p		
											0.001	0.01	0.1
UACC457 UACC383 UACC1022 TC_1376_3 TD_1376_3 TD_1730 TD_1638 TD_1720	1	-1	1	-1	-1	-1	1	1	1	-1	1	1	5
UACC3093	0	-1	1	-1	-1	-1	1	1	1	-1	11	14	30
M92_001 UACC257	1	-1	1	0	0	-1	1	1	1	-1	12	16	35
WM1791C	0	1	0	-1	1	-1	0	0	1	1	14	45	453
UACC1097	0	0	0	0	0	0	0	0	1	0	23	84	384
UACC903	0	0	0	0	0	0	1	1	0	0	32	100	539
UACC2534	1	-1	1	-1	0	-1	1	0	1	-1	35	38	50
M93_007 UACC1273	1	-1	0	0	0	0	1	1	0	0	46	30	59
UACC1265	1	-1	1	0	0	0	1	1	1	0	61	46	48
UACC091 UACC502	1	-1	0	0	0	0	1	1	0	-1	74	47	34
TD1348	0	-1	1	-1	-1	-1	1	0	1	-1	108	103	114
UACC1012	0	1	0	-1	1	0	0	0	1	0	117	143	440
M91_054	1	-1	1	-1	0	0	1	1	0	0	127	112	125
M92_047	1	1	-1	0	0	0	0	0	0	-1	136	193	453
HA_A	0	-1	1	-1	0	0	0	0	0	-1	157	122	520
TC_F027	0	-1	1	0	-1	0	1	0	1	-1	183	165	95
UACC647	0	1	0	-1	1	0	0	0	0	0	232	307	569
UACC930	0	1	-1	-1	0	0	0	0	-1	1	250	492	1275
UACC1529	-1	0	0	0	-1	0	0	0	1	-1	2625	1924	2032
UACC827T	0	0	1	-1	0	-1	0	0	1	1	2878	2383	1939
UACC2837	0	-1	-1	0	0	0	0	0	-1	0	3938	3620	4531

line thickness relatively represents the transition probability between states. The diagrams do not show transitions during transient states but only at stationary states. These can be seen as attractors for which, once have been arrived at, the chain stay within only those states. In the diagrams, the names of genes that do not change within cycles are italicized. In the first diagram, we find only one state, and once this state is achieved, the chain remains at it (based on transition rules governing the process). This may represent a core group of genes and its state. Based on current set of transition rules, there exist 395 transient states that lead to this state at the end. In the second diagram, we find two different states constructing a cycle that has 297 transient states leading to it. We also find interesting that only one of ten genes are changing its state while all other nine stay.



(a) (b)

Fig. 5. The estimated histogram distribution after long run. (a) The steady-state distribution of all possible states of the chain with perturbation probability $p = 0.1$. (b) The marginal histogram distribution for each gene.

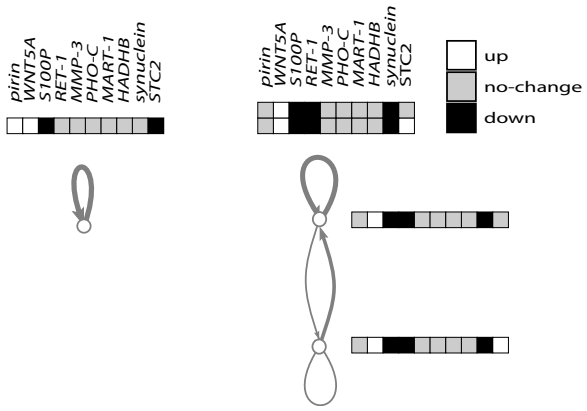


Fig. 6. State transition diagram without perturbation

For given sets of transition rules without perturbation, we often are not able to observe all possible states, thereby, missing some potentially important states. Perturbation forces the chain out of a cycle to different states. Using perturbation helps identify all non-transient states, however, after identifying these states, we may want to identify states that lead to each with high probability (its basin of attraction without perturbation), and group them, and study them as linked, but alternative pathways.

4. Conclusion

The rapidly increasing use of microarrays to develop more comprehensive views of the consequences of transcriptional regulation in cells and tissues has created a need for tools capable of producing useful inferences from this type of data. A variety of traditional mathematical and engineering tools have been adapted to this end. In this study, we examine the suitability of a mathematical tool popular in other areas such as statistics and engineering, the Markov Chain model, to describe regulatory relationships between genes. For the test, a very small network containing ten genes was built based on biological observations. The model produced steady-state distributions approximating the initial observations and exhibited many properties associated with biological systems.

The transition rules generated for the model produced localized stability. Initial states near the target states from which the model was built tended to stay in or near the target state, demonstrating that the rules were sufficient to achieve self-stabilization and to guide the system to the target state. Initial states far from the target state did not easily transition to the target state, and required assistance in the form of random perturbation to get close enough to the target state to be self-stabilizing. This requirement for close coordination of rules and contents to achieve stable states mirrors the kind of strong contextual influence seen in biology, where the rules of interaction and the state, as represented by the set of macromolecules present in the cell, are coupled in limited and characteristic ways.

The model rule sets inferred from the observations reproduced the ability of biological systems facilities in rapidly and accurately transitioning batteries of genes to very different states. The rules were sufficiently constraining to restrict the number of states seen in the steady-state, but sufficiently elastic to allow a collection of different states to be seen in the steady-state.

While the size of the problem studied in this paper is relatively small, it suggests that models incorporating rule-based transitions among states have a capacity to mimic biology. The ability of such models to enhance our understanding of biological regulation should be further tested by systematically examining the characteristics of the rules and interconnections that lead to stabilization and switch-like transitions, and by building larger networks that incorporate more extensive prior knowledge of regulatory relationships and more extensive experimental observations of the different stable states the network can occupy.

Acknowledgements

The authors wish to thank Dr. Shmulevich for his insightful suggestions on the areas of probabilistic Boolean network and Markov chain simulation.

Appendix A. Proof related to Eq. 4

Proof. The following is to prove the sum of Eq. 4 over all possible states is 1.

$$
\begin{aligned}
&\sum_{\forall \mathbf{S}^{(t+1)}} \Pr\left\{\mathbf{S}^{(t)} \to \mathbf{S}^{(t+1)}\right\} \\
&= \sum_{\forall \mathbf{S}^{(t+1)}} \left\{ \left(\prod_{l=1}^{n} C_l^{g_l^{(t+1)}} \right) \times (1-p)^n + p^{n_0}(1-p)^{n-n_0} p_0^{n_0} \times \mathbf{1}_{\left[\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}\right]} \right\} \\
&= \sum_{\forall \mathbf{S}^{(t+1)}} \left\{ \left(\prod_{l=1}^{n} C_l^{g_l^{(t+1)}} \right) \times (1-p)^n \right\} + \sum_{\forall \mathbf{S}^{(t+1)}} \left\{ p^{n_0}(1-p)^{n-n_0} p_0^{n_0} \times \mathbf{1}_{\left[\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}\right]} \right\} \\
&= (1-p)^n \sum_{\forall \mathbf{S}^{(t+1)}} \left(\prod_{l=1}^{n} C_l^{g_l^{(t+1)}} \right) + \sum_{\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}} \left\{ p^{n_0}(1-p)^{n-n_0} p_0^{n_0} \right\} \\
&= (1-p)^n \prod_{l=1}^{n} \left(\sum_{g_l^{(t+1)}} C_l^{g_l^{(t+1)}} \right) + \sum_{\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}} \left\{ p^{n_0}(1-p)^{n-n_0} p_0^{n_0} \right\} \\
&= (1-p)^n \prod_{l=1}^{n} 1 + \sum_{\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}} \left\{ p^{n_0}(1-p)^{n-n_0} p_0^{n_0} \right\} \\
&= (1-p)^n + \sum_{\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}} \left\{ p^{n_0}(1-p)^{n-n_0} p_0^{n_0} \right\}. \qquad \text{(A.1)}
\end{aligned}
$$

Since the summation in the second term is taken over all possible values of $\mathbf{S}^{(t+1)}$ except of $\mathbf{S}^{(t)} = \mathbf{S}^{(t+1)}$, n_0 can be taken from 1 to n. For a cetain n_0 bits perturbation of gene g_l which has q-nary values with equal probability p_0, there are total $\binom{n}{n_0}(q-1)^{n_0}$ states of $\mathbf{S}^{(t+1)}$. Now the second term can be rewritten as:

$$
\begin{aligned}
\sum_{\mathbf{S}^{(t)} \neq \mathbf{S}^{(t+1)}} p^{n_0}(1-p)^{n-n_0} p_0^{n_0} &= \sum_{n_0=1}^{n} \binom{n}{n_0} (q-1)^{n_0} p^{n_0} (1-p)^{n-n_0} p_0^{n_0} \\
&= \sum_{n_0=1}^{n} \binom{n}{n_0} (q-1)^{n_0} p^{n_0} (1-p)^{n-n_0} \left(\frac{1}{q-1} \right)^{n_0} \\
&= \sum_{n_0=1}^{n} \binom{n}{n_0} p^{n_0} (1-p)^{n-n_0} = 1 - (1-p)^n. \qquad \text{(A.2)}
\end{aligned}
$$

Therefore,

$$
\sum_{\forall \mathbf{S}^{(t+1)}} \Pr\left\{\mathbf{S}^{(t)} \to \mathbf{S}^{(t+1)}\right\} = (1-p)^n + 1 - (1-p)^n = 1. \qquad \text{(A.3)}
$$

□

References

[1] de Jong H., Modeling and simulation of genetic regulatory systems: A literature review, *J Comput Biol.* **9** (2002) pp. 67–103.

[2] Durbin R., Eddy S., Krogh A., and Mitchison G., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* (Cambridge Univ. Press, New York, 1998).

[3] Lange K., *Mathematical and Statistical Methods for Genetic Analysis.* (Springer-Verlag Press, New York, 1997).

[4] Kauffman S. A., *The Origins of Order, Self-Organization and Selection in Evolution.* (Oxford University Press, New York, 1993).

[5] Kauffman S. A., Metabolic stability and epigenesis in randomly constructed genetic nets, *J. Theor. Biol.* **22** (1969) pp. 437–67.

[6] Shmulevich I., Dougherty E. R., Kim S., and Zhang W. Probabilistic boolean networks: A rule-based uncertainty model for gene regulatory networks, *Bioinformatics* **18** (2002) pp. 261–274.

[7] Shmulevich I., Dougherty E. R., and Zhang W., Gene perturbation and intervention in probabilistic boolean networks., *Bioinformatics* (in press).

[8] Bittner M., Meltzer P., Chen Y., Jiang Y., Seftor, E. Hendrix M., Radmacher M., Simon R., Yakhini Z., Ben-Dor A., Sampas N., Dougherty E., Wang E., Marincola F., Gooden C., Lueders J., Glatfelter A., Pollock P., Carpten J., Gillanders E., Leja D., Dietrich K., Beaudry C., Berens M, Alberts D., Sondak V., Hayward N., and Trent J., Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature* **406** (2000) pp. 536–540.

[9] Weeraratna A. T, Jiang Y., Hostetter G., Rosenblatt K., Duray P., Bittner M. L., and Trent J. M., Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma, *Cancer Cell* **1** (2002) pp. 279–88.

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

[11] Kim S., Dougherty E. R, Bittner M. L, Chen Y., Sivakumar K. L, Meltzer P. S., and Trent J. M. A general nonlinear framework for the analysis of gene interaction via expression array, *J Biomed Opt.* **5** (2000) pp. 411–424.

[12] Kim S., Dougherty E. R, Bittner M. L, Chen Y., Sivakumar K. L, Meltzer P. S., and Trent J. M. Multivariate measurement of gene-expression relationships, *Genomics* **67** (2000) pp. 201–209.

[13] Press W. H., Teukolsky S. A., Vetterling W. T., and Flannery B. P.. *Numerical Recipes in C*, (Cambridge University Press, New York, 2nd Ed., 1992).

[14] Robert C. P. and Casella G. *Monte Carlo Statistical Methods*, (Springer-Verlag Press, New York, 1999).

[15] Stewart W. J. *Introduction to the Numerical Solution of Markov Chains*, (Princeton University Press, Princeton, NJ, 1994).