

Systems biology

Systematic intervention of transcription for identifying network response to disease and cellular phenotypes

Huai Li and Ming Zhan*

Bioinformatics Unit, Branch of Research Resources, National Institute on Aging, NIH, Baltimore, MD 21224, USA

Received on June 14, 2005; revised on September 6, 2005; accepted on October 27, 2005

Advance Access publication November 8, 2005

Associate Editor: Satoru Miyano

ABSTRACT

Motivation: A major challenge in post-genomic research has been to understand how physiological and pathological phenotypes arise from the networks of expressed genes. Here, we addressed this issue by developing an algorithm to mimic the behavior of regulatory networks *in silico* and to identify the dynamic response to disease and changing cellular conditions.

Results: With regulatory pathway and gene expression data as input, the algorithm provides quantitative assessments of a wide range of responses, including susceptibility to disease, potential usefulness of a given drug, or consequences to such external stimuli as pharmacological interventions or caloric restriction. The algorithm is particularly amenable to the analysis of systems that are difficult to recapitulate *in vitro*, yet they may have important clinical value. The hypotheses derived from the algorithm were biologically relevant and were successfully validated via independent experiments, as illustrated here in the analysis of the leukemia-associated BCR–ABL pathway and the insulin/IGF pathway related to longevity. The algorithm correctly identified the leukemia drug target and genes important for longevity, and also provided new insights into our understanding of these two processes.

Availability: The software package is available upon request to the authors.

Contact: zhanmi@mail.nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

INTRODUCTION

In a regulatory pathway of the cell, genes or proteins interact with each other to control signal transduction and transcription. The regulatory interactions are multifaceted, including protein–DNA or protein–protein interactions, one-to-one, one-to-many, or many-to-one relationships, forward or feedback loops, etc. The dynamic activity of regulatory network is constrained by the various forms of interactions, and the network thus behaves only in certain ways and controlled manners in response to changing cellular conditions or external stimuli (Huang, 2001). This study was aimed at characterizing the dynamic behavior of regulatory pathways for the understanding of how disease or cellular phenotypes arise from the connectivity or network of genes and their products, which has been a central focus in the post-genomic research era.

Important progress has been made in understanding static and dynamic properties of regulatory networks (Guelzim *et al.*, 2002;

Segal *et al.*, 2003; Shen-Orr *et al.*, 2002; Teichmann and Babu, 2003). Learning the connectivity or relationships between genes and inferring network topology and model parameters have been examined by signal processing (Kim *et al.*, 2000), pattern recognition (Liang *et al.*, 1998), Bayesian methods (Friedman *et al.*, 2000), etc. The dynamic behavior of regulatory network has been examined by the Markov chain (Kim *et al.*, 2002) or probabilistic Boolean network (Shmulevich *et al.*, 2002a). *In silico* simulation has been important in the network analysis (de Jong, 2002; Smolen *et al.*, 2000). Gene network analyses with microarray data have been used in identifying drug-affected genes or drug targets (Imoto *et al.*, 2003; Savoie *et al.*, 2003). In the present study, we mimic the behavior of the complex system of a regulatory pathway by a series of interventions made *in silico* upon each gene and combination of genes. We introduce changes consisting in altering the transcript levels of a given gene, and then calculate how much the network activity was altered in response to a certain cellular condition. The inputs to our algorithm are experiment-specific regulatory pathways and gene expression data. The outputs are the estimated probabilities of a network transit across different cellular conditions under each transcriptional intervention. The algorithm is based on the finite-state Markov chain model, which is constructed with the gene expression profile and network topology. The probability of network transition is determined based on state-dependent multivariate conditional probabilities between gene expression levels. Prior to the mathematical simulation, the regulatory network was validated by the coefficient of determination (CoD) (Dougherty *et al.*, 2000) to ensure the context specificity to the gene expression profile under examination. Real-value gene expression data were converted to the ternary presentation to ensure a high and uniform certainty in specifying genes undergoing significant transcriptional changes across experiments. The gene regulation analysis by the algorithm presented here provides critical insight into a wide range of biological processes. Specifically, the analysis would provide answers to two questions. First, whether or how much a gene or external perturbation contributes to the behavior transition of a regulatory pathway in instances such as disease development or recovery, aging process, cell differentiation or other cellular phenomena. Second, in what specific ways is this contribution manifested. The first question is answered by the measurement of the probability of network transition from one cellular state to another under transcriptional intervention of each gene or gene combination. The second question is addressed by the type of intervention applied that leads to significant network transition, i.e. upregulation, downregulation, silent transcription or a certain intervention pattern on a

*To whom correspondence should be addressed.

combination of genes. The analysis can subsequently lead to quantitative measurements of disease susceptibility, the likelihood of successful pharmacological intervention and other information that may facilitate the identification of sensitive diagnostic biomarkers and therapeutic targets. The analysis also leads to quantitative measurements of the response of the network to external perturbation like drug treatment, caloric restriction and environmental stresses, which facilitate drug screening and inform about biological processes such as aging. This analysis is particularly valuable in its ability to simulate *in silico* the pathway behaviors, which may not be easy to recreate *in vitro*. The hypotheses subsequently derived could then be tested via independent experiments.

Here, we describe the algorithm and demonstrate its usefulness by analyzing the leukemia-related BCR-ABL protein and its regulatory network, as well as by studying the activity of the age-related insulin/IGF pathway in *Caenorhabditis elegans*. The algorithm correctly identified drug targets for the leukemia model and genes important for longevity in *C. elegans* and provided new additional insight into these two biological processes. The tool provided in this work can speed researchers' efforts to unravel the structure and function of a cell's gene regulatory mechanism and to predict those cellular actions and properties controlled by such mechanisms. Furthermore, it will facilitate the development of systematic approaches for effective preventive and therapeutic intervention in disease. The potential clinical impact is tremendous as this type of intervention analysis not only can open up a window on the biological behavior of an organism and the disease progression, but also translate into accurate diagnosis, target identification, drug development and treatment. The method we proposed has made a valuable contribution on this aspect.

METHODS AND ALGORITHMS

Figure 1 illustrates the main flow of the proposed algorithm for characterization of the dynamic behavior of regulatory pathways in response to disease or cellular phenotypes. It includes three major components: CoD validation, model construction and intervention analysis. The inputs to the algorithm are experiment-specific regulatory pathways and gene expression data. The outputs are the estimated probabilities of a network transit across different cellular conditions under each transcriptional intervention.

Assessment of context specificity of network to expression profile

The CoD (Dougherty *et al.*, 2000; Kim *et al.*, 2000) was used in validating the context specificity of a network topology to expression profiles. CoD was calculated for each gene on the known pathway topology in this study. Since given genes (referred as predictors) are used to predict the behavior of a target gene under different experimental conditions, the prediction accuracy by CoD is higher in comparison with that the predictors are not used. CoD is mathematically defined as $\theta_{\text{opt}} = (\varepsilon_0 - \varepsilon_{\text{opt}})/\varepsilon_0$, where ε_0 is the prediction error in the absence of predictor and ε_{opt} is the error for the optimal predictors. The larger the CoD value is the better fitness the context specificity of a network topology to expression profiles. For showing the significance of CoD values, we also calculated the probability of obtaining a larger CoD value at random between a target gene A and given predictor genes by randomly shuffling the expression profile of predictor genes and then determining the CoD value based on the shuffled profiles and the original profile of target gene A. We repeated this process 100 000 times; the resulting distribution was fit to a Gaussian density function. We then determined the P -value of obtaining the observed CoD between gene A and given predictor genes based on the distribution.

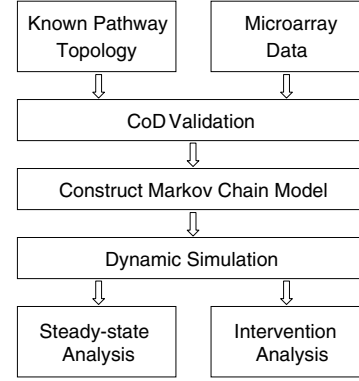


Fig. 1. Flow chart of the proposed method for the regulatory pathway validation, modeling and simulation study.

Formulation of computational model

The proposed computational model contains n selected genes. Each gene has a ternary expression value, which is assigned from over-expressed $\{1\}$, equivalently-expressed $\{0\}$, and under-expressed $\{-1\}$. For capturing the dynamics of the network, we use the state of 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 $t + 1$. Equation (1) shows the definition of transition between gene states at step t and the state at step $t + 1$, which can be represented as a Markov chain (Kim *et al.*, 2002).

$$S^{(t)} =: (g_1^{(t)} g_2^{(t)} \dots g_n^{(t)}) \rightarrow S^{(t+1)} =: (g_1^{(t+1)} g_2^{(t+1)} \dots g_n^{(t+1)}). \quad (1)$$

Here, we generalized the model which allows any number of predictor genes for each target gene based on the topology of the network. If the network topology shows there are no predictors as inputs to predict a gene in the next step, the current gene value is kept. The transition rule for $S^{(t)} \rightarrow S^{(t+1)}$ is depicted in Figure 2 and characterized by Equation (2).

$$g_l^{(t+1)} = \begin{cases} -1: & \text{with } C_l^{-1}(g_{i_1}^{(t)} g_{i_2}^{(t)} \dots g_{i_k}^{(t)}) = p(g_l^{(t+1)} = -1 | g_{i_1}^{(t)} g_{i_2}^{(t)} \dots g_{i_k}^{(t)}) \\ 0: & \text{with } C_l^0(g_{i_1}^{(t)} g_{i_2}^{(t)} \dots g_{i_k}^{(t)}) = p(g_l^{(t+1)} = 0 | g_{i_1}^{(t)} g_{i_2}^{(t)} \dots g_{i_k}^{(t)}) \\ 1: & \text{with } C_l^1(g_{i_1}^{(t)} g_{i_2}^{(t)} \dots g_{i_k}^{(t)}) = p(g_l^{(t+1)} = 1 | g_{i_1}^{(t)} g_{i_2}^{(t)} \dots g_{i_k}^{(t)}), \end{cases} \quad (2)$$

where $i_1, i_2, \dots, i_k, l \in \{1, 2, \dots, n\}$ and k is the number of predictor genes. C_l^{-1}, C_l^0 and C_l^1 are conditional probabilities that depend on the states of the predictor genes and satisfy $C_l^{-1} + C_l^0 + C_l^1 = 1$ in Equation (2). For example, if there are 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:

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

In the simulation, a small but enough perturbation is added to guarantee a steady-state distribution exists and the chain converges to the steady-state distribution. With a perturbation, the entire Markov chain is ergodic and every state will eventually be visited. Considering gene perturbation, the transition probability Equation (3) can be generalized as

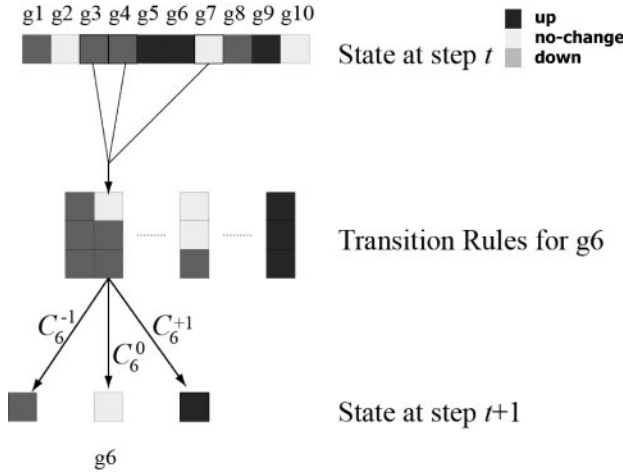


Fig. 2. Illustration of transition rules for target genes in the Markov chain model. In this example, target gene g_6 has three predictor genes g_3 , g_4 and g_7 . The value of g_6 at step $(t + 1)$ is determined by the conditional probabilities under the condition $g_3 = 0$, $g_4 = -1$ and $g_7 = -1$ at step (t) .

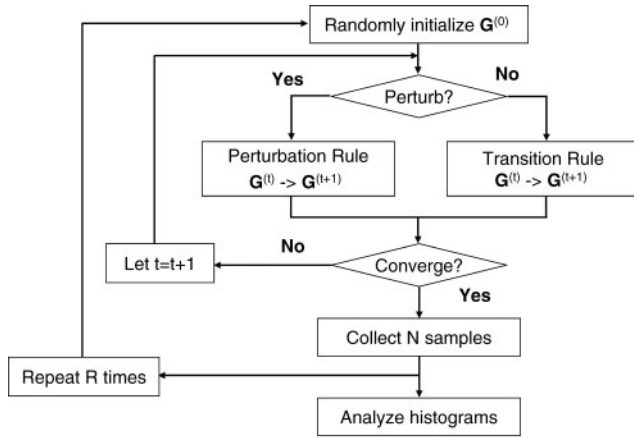


Fig. 3. Simulation algorithm for steady-state analysis. The algorithm starts from a random initial state and repeats R times before collecting samples from steady-state distribution. In the simulation, a small but measurable perturbation is added to guarantee a steady-state distribution exists and the chain converges to the steady-state distribution.

(Kim *et al.*, 2002) Equation (4):

$$\Pr\{S^{(t)} \rightarrow S^{(t+1)}\} = \left(\prod_{i=1}^n C_i^{g_i^{(t+1)}} \right) \times (1-p)^n + p^{n_0} (1-p)^{n-n_0} p_0^{n_0} \times 1_{[S^{(t)} \neq S^{(t+1)}]}, \quad (4)$$

where p is the perturbation probability for each gene, $n_0 = \sum_{i=1}^n 1_{[g_i^{(t)} \neq g_i^{(t+1)}]}$ is the number of genes to be perturbed, and $p_0 = 1/(q-1)$. In ternary case, $q = 3$, so p_0 is equal to 0.5. The simulation algorithm used this study can be summarized in Figure 3.

Intervention analysis by Markov chain model

The ability of the current model to enhance our understanding of biological regulation should be further investigated by exploring another common biological system feature, which is an ability to readily switch from one relatively stable state to another in response to a simple stimulus. To a certain

extent, this study can also verify how well the model to mimic biological systems. Basically, one question may be interesting to ask: Given a desired target state and an initial state, with which of the genes in network should we intervene genes by multiple simultaneous flipping their status so that the probability that the network will reach the desired target state is greatest?

We could address this question by finding the best candidate genes for intervention based on first-passage times (Cinlar, 1975; Shmulevich *et al.*, 2002b). The first-passage times provide a natural way to capture the goals of intervention in the sense that we wish to transit to certain states (or avoid certain states) as quickly as possible, or, alternatively, by maximizing the probability of reaching such states before a certain time. So it can be used as a tool for deciding which genes are the best candidates for intervention. The first-passage time from state x to state y can be defined as with the probability $F_k(x, y)$ that, starting in state x , the first time the network will reach a given state y will be at step k . It is easy to see that for $k = 1$, $F_1(x, y) = A(x, y)$, which is just the transition probability from x to y . For $k \geq 2$, $F_k(x, y)$ satisfies (Cinlar, 1975)

$$F_k(x, \pm y) = \sum_{z \in [-1, 0, 1]^n - \{y\}} A(x, z) F_{k-1}(z, y). \quad (5)$$

In Equation (5), each element $A(x, y)$ of the transition matrix A can be computed using Equation (4). For a fixed K , a $3^n \times K$ matrix F can be created in which each column contains the probability $F_k(x, y)$ from all possible starting states x to a given target state y at k steps. We can then use $H_K(x, y) = \sum_{k=1}^K F_k(x, y)$ as a measurement index. Because the events that the first-passage time from x to y will be at step k are disjoint for different k , the sum of their probabilities for $k = 1, \dots, K$ is equal to the probability that the network, starting in state x , will visit state y before step K . Since the chain is ergodic with perturbation probability p , when $K = \infty$, $H_\infty(x, y)$ is equal to the probability that the chain ever visits the state y , which is equal to 1.

Using the above measurement tool, we constructed the intervention information matrix H for fixing $K = 3$. In this matrix, each row $H_3(x, \cdot)$ represents the probability that the network, from a starting state x , will visit all desired ending states before step $K = 3$. Each column $H_3(\cdot, y)$ represents the probability that the network, starting in all possible intervened states, will visit state y before step $K = 3$. For simulating simple stimulus, we intervened mathematically one gene, two genes and three genes each time and kept the rest genes unchanged for a starting state x . For a ternary expression, that will generate $C_n^3 \times 3^3$ intervened states for intervening one, two and three genes, which include the original state x .

Software and experimental validation

We implemented a Java-based interactive computational tool, PathwayPro, for the algorithm that we developed (supplementary document for software introduction). With the software, we analyzed the leukemia-related ABL-BCR pathway in human and aging-related insulin/IGF-1 pathway in *C. elegans* for experimental validation of the algorithm developed. For the BCR-ABL pathway study, we used the Affymetrix array data of chronic myeloid leukemia (CML) and normal white blood cells (Crossman *et al.*, 2005; Stegmaier *et al.*, 2004), which were downloaded from the GEO database (accession numbers GSE2535 and GSE 995). For the insulin/IGF pathway study, we used the Affymetrix array data of the long-lived *daf-2* mutant and wild type of *C. elegans* (McElwee *et al.*, 2004), downloaded from the GEO database (GSE 1762). The Affymetrix microarray data were normalized by the robust multi-array analysis (RMA) method (Irizarry *et al.*, 2003).

For all datasets, we discretized gene expression values into three categories: over-expressed {1}, equivalently-expressed {0}, and under-expressed {-1} depending whether the expression level is significantly lower than, similar to or greater than the respective control threshold. Since some genes have small natural range of variation, we used z-transform to normalize the expression of genes across experiments, so that the relative expression of all genes have the same mean and standard derivation. Then we set the control threshold as 1 SD for the discretizing process.

RESULTS AND DISCUSSION

We analyzed the leukemia-related BCR-ABL pathway and the aging-related insulin/IGF pathway in *C. elegans*, to assess how effectively our algorithm profiled the network dynamic behavior in response to the leukemia development or recovery and longevity, and for the drug target discovery. The pathways were first validated against the microarray data for context specificity. *In silico* simulation was conducted by transcriptional intervention on each gene (referred to as single-gene intervention), each combination of two genes (double-gene intervention) and each combination of three genes (triple-gene intervention). In each intervention, the observed expression of a gene was altered to the opposite direction or remained unchanged. The network response was measured by the probability of network activity transit from one cellular state to another (e.g. from the normal life span to longevity state, or from leukemia to a normal condition). All computations were conducted using the software PathwayPro.

Context specificity of network to expression profile

The context of a genetic network refers to a certain state under which a limited number of genes are tightly regulated by each other via specific cellular mechanisms to perform a specific task (Huang and Ingber, 2000; Kim *et al.*, 2002). The specific task can be a different developmental stage, a tissue-specific function, a specific cell type or a certain disease. Changes in the biological context will result in changes in the abundance of sets of genes which are critically responsible for a given phenotype, and possibly also their connectivity and relationships. The set of genes tightly regulated in a specific cellular process and cell types reflecting this process can be considered as a context-specific regulatory modular network. For example, in models of cancer progression, the regulatory machinery undergoes adjustment away from the normal state to process proliferative signals and achieve a new regulatory state (Ho and Liao, 2002; Kim *et al.*, 2002). Under a certain cell type, localized similarity in gene expression patterns should be reflected in microarray data for a regulatory module. It is thus necessary to examine the context of a regulatory network and its particular gene expression profile before analyzing the associated networks.

We used the CoD measurement to assess the context specificity of pathways and validated the fitness of experiment specific microarray data to the network topology. CoD has been historically used to measure the effect of linear regression (Walpole and Myers, 1985) and for nonlinear signal processing (Dougherty *et al.*, 2000). Since CoD can treat multivariate gene relations and can discover strongly nonlinear relationships, it has been used recently in measuring multivariate interaction among genes based on gene expression (Dougherty *et al.*, 2000; Kim *et al.*, 2000), constructing probabilistic Boolean networks (Kim *et al.*, 2002; Shmulevich *et al.*, 2002a), and as an objective function for growing a subnetwork from seed genes (Hashimoto *et al.*, 2004). In this study, we viewed a regulatory pathway topology as a directed graph from predictors to targets and determined CoD to measure the strength of connection from a set of predictor genes to a target gene on the topology. The CoD value varies between 0 and 1. The larger the CoD value is, the stronger the connection is of the network. As shown in Table 1, for the insulin/IGF pathway, all CoD values of target genes are larger than 0.71 ($P \leq 0.062$), with an average of 0.875. For the ABL-BCR pathway, all CoD values of

Table 1. Evaluation of context specificity of network topology to gene expression profile by CoD

Predictor 1	Predictor 2	Predictor 3	Target	CoD	P-value
(A) Insulin/IGF pathway of <i>C. elegans</i>					
			Ins-18	N/A	N/A
			Daf-28	N/A	N/A
Ins-18	Daf-28		Daf-2	0.888	0.051
Daf-2			Ist-1	0.769	0.0075
Daf-2			Age-1	0.707	0.062
			Daf-18	N/A	N/A
Ist-1	Age-1		Pdk-1	0.920	0.060
Age-1	Daf-18	Pdk-1	Akt-1	0.983	0.034
Ist-1	Akt-1		Daf-16	0.948	0.0068
Daf-16			Sod-3	0.913	1.10E-04
(B) BCR-ABL pathway of human					
			ABL	0.742	0.036
			BCR	0.660	0.043
	ABL		CRKL	0.897	0.047
			PI3K	0.714	0.031
			PI3K	0.614	0.089
	ABL		BCR	0.890	0.070
			AKT	0.679	0.090
			AKT	0.682	0.012
	STAT5A		JAK2	0.931	0.0037
	STAT1		JAK2	0.831	0.10

(A) insulin/IGF pathway of *C. elegans*, up to 3-predictor for each target gene were evaluated based on the pathway topology. (B) BCR-ABL pathway of human, up to 2-predictor for each target gene were evaluated. The CoD value varies between 0 and 1. The value is not applicable if no predictor for a target gene.

target genes are larger than 0.66 ($P \leq 0.10$), with the average of 0.764. The results are indicative of strong connectivity of genes on the network with high significance based on the gene expression data utilized. The pathway topology is therefore context specific to the expression data, and the expression data are suitable for the analysis of the two pathways.

Disease gene and drug target of leukemia

Figure 4A shows the network topology of the ABL-BCR pathway (Lugo *et al.*, 1990; Raitano *et al.*, 1997; Zou and Calame, 1999). BCR and ABL are linked to the cytoplasm as a part of a large signaling complex with a variety of cellular substrates, related to the development of CML (Lugo *et al.*, 1990; Raitano *et al.*, 1997; Zou and Calame, 1999). We measured the transition probability of the ABL-BCR pathway between the normal condition and leukemia state under a series of transcriptional interventions. The probability of the network transitioning from normal to leukemia reveals disease susceptibility of genes involved. The higher the probability is, the more likely a gene or gene combination under a certain intervention is responsible for the development of the disease. On the other hand, the probability of the transition from leukemia to normal states is a measure of the potential usefulness of a drug or therapeutic intervention.

Our analysis first indicates that more genes and gene combinations have higher probabilities to contribute to regulatory network transitions from normal to leukemia than from leukemia to normal, at a certain threshold (Tables 2 and 3). It therefore suggests that the chance is higher for human to develop leukemia than to recover

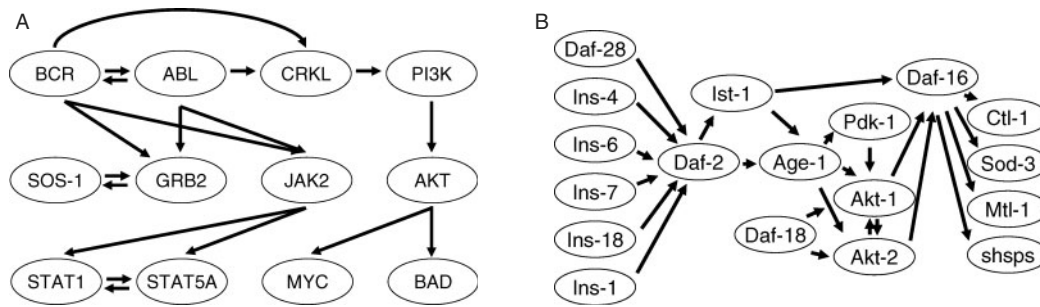


Fig. 4. (A) The leukemia-related BCR-ABL pathway and the inhibitory effect of the drug Gleevec. (B) The aging-related insulin/IGF pathway of *C. elegans*.

Table 2. Probabilities of network transition under double-gene interventions on genes in the ABL-BCR associated pathway of human

Genes	Transcriptional intervention	Transition probability
(A) Transition from the normal to CML states		
BCR ABL1	0 -1 ⇒ 1 1 ⇒ 1 1	0.010933
BCR BAD	0 1 ⇒ -1 0 ⇒ 1 0	0.006388
BCR MYC	0 -1 ⇒ -1 0 ⇒ 1 0	0.006388
BCR BAD	0 1 ⇒ -1 -1 ⇒ 1 0	0.006388
BCR MYC	0 -1 ⇒ -1 1 ⇒ 1 0	0.006388
BCR STAT5A	0 1 ⇒ -1 -1 ⇒ 1 1	0.006388
BCR STAT5A	0 1 ⇒ -1 0 ⇒ 1 1	0.006388
BCR STAT1	0 0 ⇒ -1 1 ⇒ 1 0	0.006388
BCR STAT1	0 0 ⇒ -1 -1 ⇒ 1 0	0.006388
BCR CRKL	0 -1 ⇒ -1 1 ⇒ 1 0	0.005367
BCR CRKL	0 -1 ⇒ -1 0 ⇒ 1 0	0.003978
BCR PIK3CG	0 -1 ⇒ -1 0 ⇒ 1 -1	0.003836
BCR JAK2	0 0 ⇒ -1 1 ⇒ 1 0	0.002239
BCR AKT1	0 0 ⇒ -1 -1 ⇒ 1 0	0.001066
(B) Transition from the CML to normal states		
ABL1 AKT1	1 0 ⇒ 0 1 ⇒ -1 0	0.001850436
ABL1 AKT1	1 0 ⇒ 0 -1 ⇒ -1 0	0.001791257
BCR ABL1	1 1 ⇒ 0 -1 ⇒ 0 -1	0.00110597

The probability cutoff is 0.001. (A) Transition from the normal to CML states; (B) Transition from the CML to normal states. Transcriptional intervention is presented as initial state (e.g. normal state) ⇒ state after intervened ⇒ end state (e.g. disease state). In each state, expression levels of each gene are presented by ternary values.

from the disease. As shown in Table 2, in the double-gene intervention, changes directly involving the genes BCR and ABL yield the highest probability (0.01) for a normal-to-leukemia transition. The interventions on ABL/AKT1 and BCR/ABL lead to the highest transition probabilities (0.002 and 0.001, respectively) for a leukemia-to-normal transition, although they remain nearly 100 times lower than those for normal-to-leukemia transitions. In the triple-gene intervention (Table 3), the triplets BCR/ABL/BAD and BCR/ABL/MYC show the highest probability (0.01) for normal-to-leukemia transition, while the BCR/ABL/AKT combination appears to have the highest probability (0.007) for leukemia-to-normal transitions. The importance of BCR and ABL to the network transition is further illustrated by the single-gene intervention, where the two genes are associated with the highest transition probability of (Table 3). Moreover, BCR and ABL show high frequencies in all of their partnerships with other genes in the double or triple interventions positive for network transition. As shown in

Table 3. Probabilities of network transition by serial interventions on genes in the ABL-BCR pathway of human

Gene	Transcriptional intervention	Transition probability
(A) Transition from normal to CML states by single-gene intervention ^a		
BCR	0 ⇒ -1 ⇒ 1	0.006387523
(B) Transition from CML to normal states by single-gene intervention ^b		
ABL1	1 ⇒ 0 ⇒ -1	0.000298692
(C) Transition from normal to CML states by triple-gene intervention ^c		
BCR ABL1 BAD	0 -1 1 ⇒ 1 1 0 ⇒ 1 1 0	0.010936278
BCR ABL1 MYC	0 -1 -1 ⇒ 1 1 0 ⇒ 1 1 0	0.010936278
BCR ABL1 BAD	0 -1 1 ⇒ 1 1 -1 ⇒ 1 1 0	0.010933351
BCR ABL1 MYC	0 -1 -1 ⇒ 1 1 1 ⇒ 1 1 0	0.010933351
BCR ABL1 STAT5A	0 -1 1 ⇒ 1 1 0 ⇒ 1 1 1	0.010933348
BCR ABL1 STAT5A	0 -1 1 ⇒ 1 1 -1 ⇒ 1 1 1	0.010933348
BCR ABL1 STAT1	0 -1 0 ⇒ 1 1 -1 ⇒ 1 1 0	0.010933348
BCR ABL1 STAT1	0 -1 0 ⇒ 1 1 1 ⇒ 1 1 0	0.010933348
(D) Transition from CML to normal states by triple-gene intervention ^d		
BCR ABL1 AKT1	1 1 0 ⇒ 0 -1 1 ⇒ 0 -1 0	0.006842
BCR ABL1 AKT1	1 1 0 ⇒ 0 -1 -1 ⇒ 0 -1 0	0.006624
ABL1 CRKL AKT1	1 0 0 ⇒ 0 -1 1 ⇒ -1 -1 0	0.002973
ABL1 CRKL AKT1	1 0 0 ⇒ 0 -1 -1 ⇒ -1 -1 0	0.002878
BCR ABL1 AKT1	1 1 0 ⇒ -1 -1 1 ⇒ 0 -1 0	0.002741
BCR ABL1 AKT1	1 1 0 ⇒ -1 -1 -1 ⇒ 0 -1 0	0.002653
ABL1 CRKL AKT1	1 0 0 ⇒ 0 1 1 ⇒ -1 -1 0	0.002498
ABL1 CRKL AKT1	1 0 0 ⇒ 0 1 -1 ⇒ -1 -1 0	0.002418

The gene expression profile of each state is presented as initial state (e.g. normal state) ⇒ state after intervened ⇒ end state (e.g. disease state). Transcriptional intervention is presented as initial state (e.g. normal state) ⇒ state after intervened ⇒ end state (e.g. disease state). In each state, expression levels of each gene are presented by ternary values.

Note: Transition probabilities by double-gene interventions are listed in Table 2 of the paper.

^aProbability cutoff 1E - 4.

^bProbability cutoff 1E - 4.

^cProbability cutoff 1E - 2.

^dProbability cutoff 2E - 3.

Figure 5, BCR and ABL are on the top by the frequency of partnership with other genes in the normal to leukemia transition, while BCR and ABL, along with AKT and CRKL, are on the top in the leukemia to normal transition in the triple-gene invention. A similar situation was also observed in the double-gene intervention (Figure 6). It can therefore be concluded that BCR and ABL are the most contributive to the network behavior transition between the normal condition and the leukemia state, and therefore the most susceptible for the development of the CML as well as the recovery

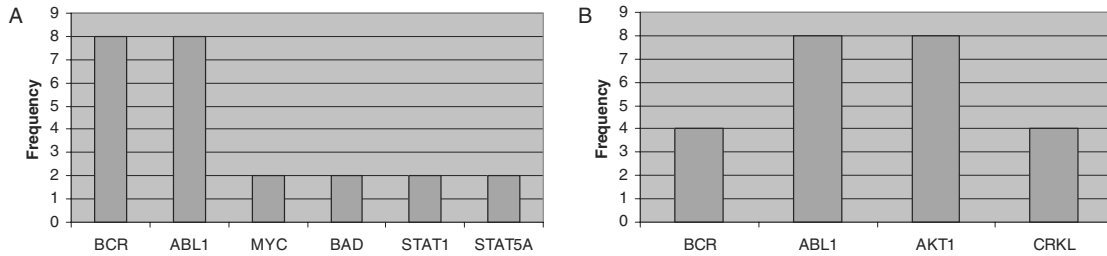


Fig. 5. Frequency of partnership of each gene with other genes in the positive triple-gene interventions on the ABL-BCR associated pathway. (A) Transition from normal to CML states (transition probability cutoff: 0.01); (B) Transition from CML to normal states (transition probability cutoff: 0.001).

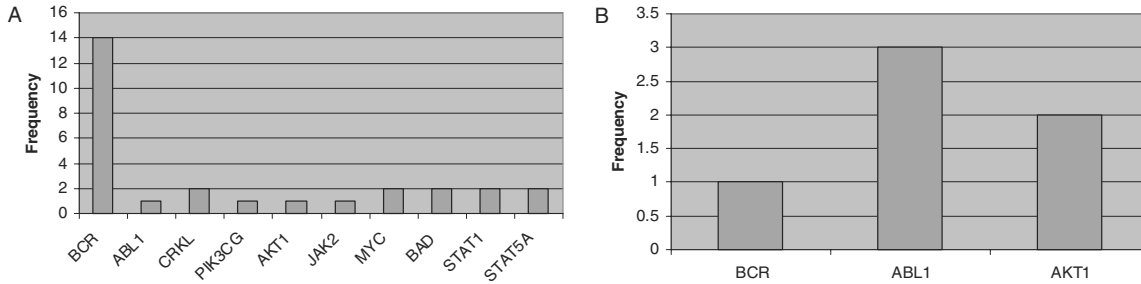


Fig. 6. Frequency of partnership of each with other genes in positive gene interventions on ABL-BCR pathway. (A) Transition from normal to CML states by double interventions (probability cutoff $1E-3$); (B) Transition from CML to normal states by triple interventions (probability cutoff $1E-3$).

of the disease to a normal condition, based on the regulatory pathway behavior. The two genes can thus serve as good drug target candidates for the treatment of the CML. This result, reached independently by the computational analysis, is in agreement with the conclusion by previous laboratory-based studies. It has been shown that CML is associated in most cases with the fusion of the genes ABL and BCR, and the activation of BCR-ABL represses apoptosis and allows transformed cells to divide, resulting in the development of CML (Lugo *et al.*, 1990; Raitano *et al.*, 1997; Zou and Calame, 1999). The BCR-ABL genes are therefore considered as ideal drug targets, and the drug Gleevec is a selective BCR-ABL inhibitor, effective in the treatment of CML (Druker *et al.*, 2001). Our computational analysis not only correctly identified the drug target, but also further indicated that BAD and MYC play critical role in the leukemia development while AKT appears important in the leukemia recovery to normal, providing new insights into our understanding of the disease.

Longevity and aging in *C. elegans*

Longevity and aging in *C. elegans* is controlled in part by the insulin/insulin-like growth factor-1 signaling pathway (Gottlieb and Ruvkun, 1994; Patterson and Padgett, 2000). The pathway is evolutionarily conserved and therefore studies of insulin-like control of longevity in *C. elegans* would help to reveal mechanisms that control human longevity (Gottlieb and Ruvkun, 1994; Patterson and Padgett, 2000). Figure 4B shows the topology of the insulin/IGF-1 pathway in *C. elegans* (Gottlieb and Ruvkun, 1994; Patterson and Padgett, 2000). Daf-2 is a component gene of the pathway, and the daf-2 mutant is similar to the long-lived dauer larva in both life span and gene expression profile and represents known and potential determinants of longevity (Liu *et al.*, 2004). We therefore used the gene expression data from the daf-2 mutant in comparison

with the wild type to identify dynamics behavior of the pathway in response to longevity and aging in *C. elegans*.

Our analysis results are summarized in the Supplementary Table 1. We first found that more genes or gene combinations had higher probabilities to contribute to network transitions from a longevity state to normal life span than that from normal life span to longevity. As shown, for a longevity-to-normal transition, 8 triple combinations of totally 7 genes are contributive at the threshold of 0.02, with top transition probabilities observed in the triplet *ist-1/pdk-1/daf-16* (probability 0.1), followed by *pdk-1/akt-1/daf-16* and *ist-1/pdk-1/akt-1* (0.06). For a normal-to-longevity transition, however, only 4 triple combinations of totally 5 genes reach the same threshold, with top probabilities observed for *pdk-1/akt-1/daf-16* (0.074) and *age-1/akt-1/daf-16* (0.041). Similar trends were also observed by the double-gene intervention. These data suggest that it is less likely for the animal to live longer than to have a normal life span, according to the dynamic behavior of the insulin/IGF-1 pathway. As we can see, the genes *daf-16*, *akt-1*, *akt-2*, *pdk-1*, and *ist-1* are highly contributive to the network transition between normal life span and longevity state, which was detected not only by the triple interventions, but also by double and single interventions, although the transition probabilities are relatively low in the latter two cases. In the double intervention, the gene doublets *pdk-1/akt-1* and *pdk-1/ist-1* are associated with high probabilities in the network transition from longevity to normal states. The transition probability of *pdk-1/ist-1* is 0.03 when *ist-1* expression is lowered and *pdk-1* remains not changed, while it drops to 0.024 when the expression of the two genes is reduced. The transition probability of *pdk-1/akt-1* is 0.03 when *akt-1* levels are reduced and *pdk-1* remains unchanged, while drops to 0.015 when two genes are both reduced. It seems important to maintain constant levels of *pdk-1* in order for transition from longevity to normal states. The importance of *daf-16*, *akt-1*,

akt-2, pdk-1 and ist-1 for the network transition between the normal life span and longevity, as suggested by the transition probability, is underscored by the high frequency of each of the genes in partnering with other genes in all positive double or triple interventions (Supplementary Figure 1). In the normal life span-to-longevity transition, daf-16 is the most frequent gene detected, followed by pdk-1. In the longevity-to-normal life span transition, pdk-1 is the most frequent, followed by daf-16 or ist-1. It can thus be concluded that daf-16 is a major contributor for the network transition from the normal life span to longevity, while pdk-1 a major contributor for the transition from longevity to the normal life span. This conclusion is consistent with the result by many previous laboratory-based studies (Kenyon, 1993; Lin, 2001). It has been shown that the longevity of daf-2 mutants is dependent on daf-16. Daf-16 in daf-2 mutants accumulates in the nuclei of many cell types and leads to changes in the expression of a wide variety of genes including those affecting metabolism, the stress response and antimicrobial functions, and thereby extends lifespan (Lin, 2001; Murphy, 2003). Our computational analysis further suggests a unique role of ist-1 in the aging process, which was not identified in previous studies.

A major challenge in post-genomic research has been to understand how phenotypes influencing disease or cellular phenotypes arise from regulatory gene networks. By modeling the behavior of regulatory pathways and gene connections in specific biological settings, the algorithm described here allows the elucidation of how much and in what ways genes or external perturbations contribute to disease development, the aging process and other cellular phenomena. The algorithm provides a new tool for drug discovery, for the identification of sensitive diagnostic biomarkers, as well as for basic investigation of many aspects of systems biology.

ACKNOWLEDGEMENTS

The authors wish to thank Drs M. Gorospe, M. Rao, S. Zou, M. Ko and X. Xu for insightful suggestions and helpful discussions. We also thank the anonymous reviewers for the critical comments.

Conflict of Interest: none declared.

REFERENCES

- Cinlar,E. (1975) *Introduction to Stochastic Processes*. Prentice Hall, NJ.
- Crossman,L.C. *et al.* (2005) In chronic myeloid leukemia white cells from cytogenetic responders and non-responders to imatinib have very similar gene expression signatures. *Haematologica*, **90**, 459–464.
- de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.
- Dougherty,E.R. *et al.* (2000) Coefficient of determination in nonlinear signal processing. *Signal Processing*, **80**, 2219–2235.
- Druker,B.J. *et al.* (2001) Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N. Engl. J. Med.*, **344**, 1038–1042.
- Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
- Gottlieb,S. and Ruvkun,G. (1994) daf-2, daf-16 and daf-23: genetically interacting genes controlling dauer formation in *Caenorhabditis elegans*. *Genetics*, **137**, 107–120.
- Guelzim,N. *et al.* (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.*, **31**, 60–63.
- Hashimoto,R. *et al.* (2004) Growing genetic regulatory networks from seed genes. *Bioinformatics*, **20**, 1241–1247.
- Ho,K.J. and Liao,J.K. (2002) Non-nuclear actions of estrogen: new targets for prevention and treatment of cardiovascular disease. *Mol. Interv.*, **2**, 219–228.
- Huang,S. (2001) Genomics, complexity and drug discovery: insights from Boolean network models of cellular regulation. *Pharmacogenomics*, **2**, 203–222.
- Huang,S. and Ingber,D.E. (2000) Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell. Res.*, **261**, 91–103.
- Imoto,S. *et al.* (2003) Use of gene networks for identifying and validating drug targets. *J. Bioinform. Comput. Biol.*, **1**, 459–474.
- Irizarry,R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.
- Kenyon,C. *et al.* (1993) A *C. elegans* mutant that lives twice as long as wild type. *Nature*, **366**, 461–464.
- Kim,S. *et al.* (2000) A general nonlinear framework for the analysis of gene interaction via expression array. *J. Biomed. Optics*, **5**, 411–424.
- Kim,S. *et al.* (2002) Can Markov chain models mimic biological regulation? *J. Biol. Syst.*, **10**, 337–357.
- Liang,S. *et al.* (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, 18–29.
- Lin,K. *et al.* (2001) Regulation of the *Caenorhabditis elegans* longevity protein DAF-16 by insulin/IGF-1 and germline signaling. *Nat. Genet.*, **28**, 139–145.
- Liu,T. *et al.* (2004) Regulation of signaling genes by TGF β during entry into dauer diapause in *C. elegans*. *BMC Dev. Biol.*, **4**, 1–17.
- Lugo,T.G. *et al.* (1990) Tyrosine kinase activity and transformation potency of BCR-ABL oncogene products. *Science*, **247**, 1079–1082.
- McElwee,J.J. *et al.* (2004) Shared transcriptional signature in *Caenorhabditis elegans* Dauer larvae and long-lived daf-2 mutants implicates detoxification system in longevity assurance. *J. Biol. Chem.*, **279**, 44533–44543.
- Murphy,C.T.M. *et al.* (2003) Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature*, **424**, 277–283.
- Patterson,G.I. and Padgett,R.W. (2000) TGF- β -related pathways: roles in *C. elegans* development. *Trends Genet.*, **16**, 27–33.
- Raitano,A.B. *et al.* (1997) Signal transduction by wild-type and leukemogenic Abl proteins. *Biochim. Biophys. Acta*, **1333**, 201–216.
- Savoie,C.J. *et al.* (2003) Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. *DNA Res.*, **10**, 19–25.
- Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.
- Shen-Orr,S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**, 64–68.
- Shmulevich,I. *et al.* (2002a) Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**, 261–274.
- Shmulevich,I. *et al.* (2002b) Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics*, **18**, 1319–1331.
- Smolen,P. *et al.* (2000) Modeling transcriptional control in gene networks—methods, recent results, and future directions. *Bull. Math. Biol.*, **62**, 247–292.
- Stegmaier,K. *et al.* (2004) Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nat. Genet.*, **36**, 257–263.
- Teichmann,S.A. and Babu,M.M. (2003) Gene regulatory network growth by duplication. *Nat. Genet.*, **36**, 492–496.
- Walpole,R.E. and Myers,R.H. (1985) *Probability and Statistics for Engineers and Scientists*. Macmillan, NY.
- Zou,X. and Calame,K. (1999) Signaling pathways activated by oncogenic forms of Abl tyrosine kinase. *J. Biol. Chem.*, **274**, 18141–18144.