Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 246 (2007) 234–244
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Cancer onset and progression: A genome-wide, nonlinear dynamical systems perspective on onconetworks

K. Qu, A. Abi Haidar, J. Fan, L. Ensman, K. Tuncay, M. Jolly, P. Ortoleva*

Center for Cell and Virus Theory, Department of Chemistry, Indiana University, Bloomington, IN 47405–7102, USA

Received 9 March 2006; received in revised form 21 November 2006; accepted 1 December 2006
Available online 15 December 2006

Abstract

It is hypothesized that the many human cell types corresponding to multiple states is supported by an underlying nonlinear dynamical system (NDS) of transcriptional regulatory network (TRN) processes. This hypothesis is validated for epithelial cells whose TRN is found to support an extremely complex array of states that we term a "bifurcation nexus", for which we introduce a quantitative measure of complexity. The TRN used is constructed and analyzed by integrating a database of TRN information, cDNA microarray data analyzers, bioinformatics modules, a transcription/translation/post-translation kinetic model, and NDS analysis software.

Results of this genome-wide approach suggest that a cell can be induced to persist in one state or to transition between distinct states; apparently irreversible transitions can be reversed when the high dimensional space of extracellular and intracellular parameters is understood. As conditions change, certain cellular states (cell lines) are no longer supported, new ones emerge, and transitions (cell differentiation or death) occur. The accumulation of simulated point mutations (minor changes which individually are insignificant) lead to occasional dramatic transitions. The genome-wide scope of many of these transitions is shown to arise from the cross-linked TRN structure. These notions imply that studying individual oncogenes may not be sufficient to understand cancer; rather, "onconetworks" (subsets of strongly coupled genes supporting multiple cell states) should be considered. Our approach reveals several epithelial onconetworks, each involving oncogenes and anti-tumor and supporting genes.

Keywords: Bifurcation nexus; Cell differentiation; Onconetwork; Microarray data; Gene ontology; Nonlinear dynamical system; Regulatory network

1. Introduction

Classic work in cancer genomics has focused on the discovery of oncogenes and tumor suppressor genes (Sassone-Crosi et al., 1988 and Table 1). These are the elements on which much of our understanding of cancer is built. The importance of oncogenes as initiators of directional signaling pathways has been suggested (Vogt et al., 1999). Models of small sub-networks constructed around some of these "key" genes have yielded insights into cancer onset and progression (Obeyesekere et al., 2004; Hervagault et al., 1991). However, cell transformation can be genome-wide in scope due to extensive gene–gene cross-linking in the structure of the human transcription regulatory network (TRN), and these restricted models overlook many important effects, as evidenced by unforeseen drug side-effects and acquired drug resistance. It is even more difficult for these models to answer fundamental questions such as why the transition to cancer may occur without dramatic changes in chromosomal sequence; why microbes make dramatic yet reversible changes in metabolism or physiology in response to environmental variations; and why multi-cellular organisms develop a myriad of differentiated cell types displaying major differences in cell behavior with the same DNA sequence. In this paper, we propose that these diverse phenomena can be described using one model, and we will demonstrate this using a genome-wide human TRN.

Equations describing cellular reaction-transport processes are nonlinear in concentrations, membrane potentials, etc., therefore, a cell can be considered to be a nonlinear dynamical system (NDS). Attempts to analyze cellular NDS problems date back to Turing (1952) who

*Corresponding author. Tel.: +1 812 855 2717; fax: +1 812 855 8300.
E-mail address: ortoleva@indiana.edu (P. Ortoleva).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.12.002

Table 1
Most important genes (up to 35) for each of the four zones seen in Fig. 3

	Zone 1		*Zone* 2		*Zone* 3		*Zone* 4	
1	STAT1	44	MYBL1*	20	NFE2L1	807	HIF1A	1693
2	TP73^{+}	27	SMAD3	5	FOXO3A	546	NFE2L1	1609
3	FOXO3A	25	POU2AF1	3	TBP	285	STAT3	311
4	NR3C1	25	NR3C1	3	GATA1	267	GLUR	294
5	BRCA1^{+}	25	POU2F2	2	ATF4	190	TBP	285
6	TP53^{+}	24	CEBPB	1	NR3C1	143	STAT5B	163
7	STAT5B	19	ZNF148	1	USF2	139	STAT5A	147
8	RELA	18	TFEB	1	JUND	132	STAT2	147
9	NFKB1	17	BRCA1^{+}	1	DBP	132	CEBPD	83
10	STAT2	16			TCF4	127	USF2	81
11	STAT5A	16			TBX2	108	JUND	79
12	P63	14			KLF2	105	TCF4	79
13	TNFRSF25	14			JUNB	83	TBX2	73
14	ZNF148	13			ZNF148	82	FOXO3A	70
15	POU2AF1	10			NFKB1	68	KLF2	65
16	FOSL2^{+}	8			FOSL1^{+}	59	JUNB	55
17	RBL2	8			RUNX1	49	ZNF148	42
18	PAX3	7			TFEB	45	NR3C1	30
19	JUN*	6			MYBL1*	33	FOXA3	23
20	ZNFN1A1	6			FOXA3	32	RUNX1	23
21	SP4	6			BRCA1^{+}	25	NFKB1	21
22	RXRA	6			ELF1	23	FOSL1^{+}	18
23	g_$^{+}$)	6			NFYB	23	BRCA1^{+}	14
24	g_-)	6			FOXA1	21	RXRA	11
25	g_AGE2	6			RXRA	17	MYC*	10
26	g_AP-3_(1)	6			TFAP4	13	ELF1	9
27	g_ATF-6	6			VDR	12	TFEB	9
28	g_BAP	6			ATF1	10	NFYB	9
29	g_BCL-9	6			YY1	10	MYBL1*	9
30	g_BRCA2^{+}	6			SPI1*	9	SRF	9
31	g_CARG	6			STAT1	9	FOXA1	8
32	g_CBF1	6			PAX3	8	DBP	8
33	g_CDK4^{+}	6			PGR	8	ATF4	7
34	g_CLIF	6			HSF1	7	VDR	6
35	g_COUP-TF1	6			TP53^{+}	7	TFAP4	6

The importance factor is the shaded area shown in Fig. 4 as described in Section 2. Here g_xyz represents the gene that encodes protein xyz. Oncogenes (*) and tumor suppressor genes (+) are from http://embryology.med.unsw.edu.au/DNA/DNA10.htm.

proposed the concept of self-organized patterns in interacting cells. Reshevsky emphasized the role of nonlinearity in biological systems (Rashevsky, 1960) while Kauffman studied multiple cellular states based on a boolean model (Kauffman, 1969); these authors suggested that the distinct steady states supported by an NDS might be distinguishable cell lines or types (some normal, some cancerous), and transitions between these states are associated cell transformations. Later Prigogine and coworkers systematically described self-organization, demonstrating the importance of far-from-equilibrium conditions in supporting NDS phenomena (Nicolis and Prigogine, 1977). Recent research has focused on cell differentiation (Ortoleva and Ross, 1973a, b), cell modeling (Ortoleva et al, 2003; Novak and Tyson, 1993), and multiple steady-states analysis (Hanns-gen and Tyson, 1985; Mochizuki, 2005). Presently, there is a great interest in delineation gene regulatory networks and deriving their implications for cell differentiation based on the integration of genomic, proteomic, metabolic and other data. Our objective here is to relate the biological phenomena of distinct cell types to the properties of the TRN and the NDS effects it supports.

Transitions associated with nonlinear cellular dynamics are frequently assumed to be caused by mutations that change the structure of the TRN. For example, cancer may occur after months or years, leaving the impression that it is a result of such a rare dramatic mutation. Yet minor mutations at the local sequence level may leave the structure of the TRN intact, instead altering, for example, transcription factor (TF)/gene binding or transcription rate constants. In the cell system, a minor sequence change can move a key physicochemical parameter, but no dramatic change in behavior will occur until a critical state is reached. At this point, a dramatic transition in RNA expression and other key cell activity is suddenly triggered (without changing the basic structure of the TRN). In this process, no individual point mutation is actually responsible; rather, it is a cumulative effect. It may take years for a cell line to accumulate enough point mutations to pass the critical point where the cellular NDS transitions to abnormal cell behavior.

The notions suggested above can also be applied to cases wherein the abnormal behavior is induced by, for example, the presence of an excessive number of copies of a given gene. Extra gene copies would effectively increase the transcriptional rate constant for that gene, and this might move the cell closer to or beyond the point of transition to abnormality. Thus, patients with an abnormal number of copies of a given gene may be more vulnerable to the onset of cancer. This might be the case for erbB2 (also known as HER2), a gene that is overabundant in 15–25% of breast cancer patients (Piccart-Gebhart et al., 2005). Our long term goal is to show how transitions among states of the cellular NDS can be related to the onset and progression of cancer in this and similar situations.

It may be argued that in many cases, cell transformations are irreversible; however, we have shown (Hahn et al., 1973) that the apparent irreversibility of transitions among states of an NDS can be overcome. While the system may not be able to return to a previous steady state by changing only the parameter that caused the system to leave that steady state, nevertheless, changes in one or more other parameters may be able to modify the folded steady-state branch structure and overcome apparent irreversibility. It is the existence of many system parameters that allows such reversibility, and identifying them is a promising area for cancer treatment discovery.

A human cell NDS, which can be represented schematically by the triangular structure in Fig. 1(a), contains about 25 000 genes, each associated with many kinetic parameters for transcription, translation, post-translational modification, and other processes. There are thousands of physicochemical parameters needed to describe the extracellular medium as well as hundreds of thousands or more cellular state variables (e.g. local concentrations, membrane potentials, and cell physiology). These variables reside in a space of millions of dimensions and may allow reversibility. Furthermore, a real NDS is better represented as a tetrahedral structure in which the triangle of Fig. 1(a) interacts with metabolites, enzymes, and cellular architecture. This tetrahedral network interacts with chemical and thermal factors imposed by the extracellular medium, making the whole system even more complex.

We hypothesize that the incredibly complex nesting of bifurcation structure (which we term a "bifurcation nexus") is a key aspect of human cell behavior and is likely a central feature of the onset and progression of cancer. The complexity of the NDS of a human TRN supports the possibility of cell transformation, but it also make it very challenging to construct and analyze the networks to find the states supported by the human cell and to identify the nature of, and conditions inducing transitions among, these states. To overcome this difficulty, we have developed an automated strategy to discover and analyze a human cell's genome-wide network of regulatory processes. The TRN is constructed from a preliminary network consisting of known gene-TF interactions, and is greatly augmented via predictions from bioinformatics modules. Quantitative transcription/translation/post-translational process modeling was used to obtain parameters for the NDS analysis (see Section 2, Appendix A).

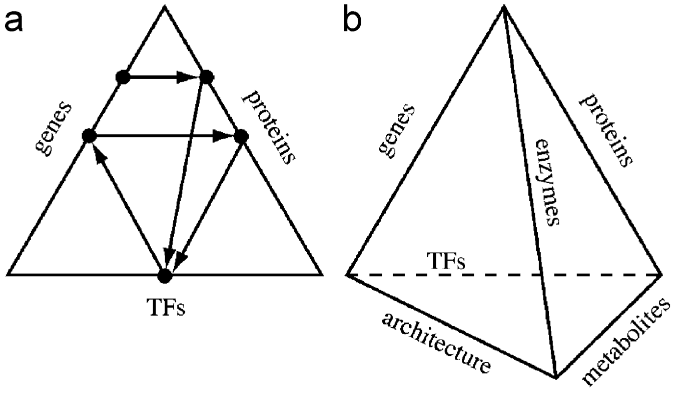

Fig. 1. (a) In the simplest model, genes encode proteins which form TFs that regulate genes. (b) In reality, enzymes, metabolites and cellular architecture are also key elements of the cell's nonlinear dynamical system, represented here as an interaction tetrahedron. The latter is embedded in a sphere (not shown) that represents extracellular medium.

In this paper, we apply this strategy to analyze a breast cancer microarray experiment (Lin et al., 2004), and demonstrate that the approach can be used to derive the cell biological implication from microarray data. The model use is based on a genome-wide network and is not crafted to arrive at a predetermined conclusion. The results of the model might be helpful in understanding RNA, TF, and protein concentrations at a quantitative level that may facilitate the design of drug cocktails (for example, the drug cocktails being investigated for certain breast cancers (Marty et al., 2005; Pietras et al., 1998; Baselga et al., 1998; Pegram et al., 1999). The automated network discovery and analysis strategy we have developed (sysbio.indiana.edu) will also benefit system biology research in general.

2. Epithelial cell demonstration

The existence of a variety of human cell types, transitions among them, and the potential relevance to cancer onset and progression are considered using the approach outlined below. The TRN used was developed via the strategy summarized in Appendix B. Here we investigate a TRN constructed for an epithelial cell to demonstrate transitions between cell states using the NDS analysis and stochastic cell line trajectory methods of Appendices A and B.

The epithelial cell TRN studied was obtained using microarray data on the estrogen response of breast cancer cells (Lin et al., 2004). We first included genes that were estrogen-responsive in the cDNA microarray data, or that were otherwise believed to play a key role in breast cancer according to the literature. Also included were the genes that wholly or partially encode the TFs that regulate the selected genes. The complete list of genes and TFs that regulate them, as well as the sense (up/down) of the regulation, is provided in Appendix C. Other data used in our analysis is summarized in Appendix D and includes estimates of rate and binding constants for transcriptional, translational and post-translational processes. The stoichiometry of these processes was obtained from our GeneDat database of experimentally verified information gathered from the literature and other public databases.

Additional features of the system investigated are now discussed. TFs were assumed to be homo-dimerized unless another type of complex was indicated in GeneDat to be the active form. The TF/gene interactions used include those in GeneDat bolstered by those predicted by our GO analysis of all human chromosomes. The GO predictions were based on biological function and the assumption that genes with high GO similarity are likely to be regulated in a similar fashion. We ran our FTF microarray data interpreter (see Appendix B) to obtain a TRN that was most consistent with the microarray data, thereby modifying some TF/gene regulatory interactions and adding others. Changing some of the TF/gene interactions obtained from GeneDat is justified as not all regulatory information used was from the same cell line. In a sense, the microarray analysis allows us to integrate regulatory information from a number of cell lines, keeping only those that are consistent with the observed expression profiles for the cell line of interest (here human epithelial cells). We decomposed the resulting TRN to eliminate decoupled subnetworks and thereby simplify the analysis by reducing the dimensionality of the problem. This network was enriched with a few post-translational reactions to investigate TF activation by phosphorylation. Our chemical kinetic cDNA microarray analyzer, KAGAN (see Appendix B), was used to estimate TF/gene binding constants and rate coefficients for transcription when we had high-quality RNA expression data for a given gene. If high quality microarray data was not available for a gene, the binding constants for all its up-regulations were set equal to the average of the binding constants of all up-regulations determined by KAGAN, and similarly for down-regulations. A matrix specifying the gene that encodes each TF component protein, and the proteins that complex to form each active TF, was extracted from GeneDat. The equations for the model of Appendices A and B were then used via an automated preprocessor to generate the computer-readable set of differential equations required by a version of the AUTO (Doedel et al., 1991a, b) NDS analysis package (modified to allow for the many variables in the model of Appendix A). Parameter values used were as discussed in Appendix D.

The erbB2 gene is overexpressed in 15–25% of breast cancers (Piccart-Gebhart et al., 2005) and thus was added to the TRN studied. The up-regulation of erbB2 by ER81 is indirectly related to the transactivation of ER81 by a positive feedback regulatory loop. In particular, HER2 (the protein encoded by erbB2) is involved in the modifications made to ER81 to render it maximally transactive. These transcription-activating modifications are the phosphorylation and acetylation of ER81 by complexes that either include or are stimulated by HER2 (15–18 Goel and Janknecht, 2003, 2004; Bosc and Janknecht, 2002; Bosc et al., 2001). The participation of the HER2 protein as a phosphorylation-promoting enzyme was investigated by adding several equations to the model. For example, one equation describes erbB2 regulated by several TFs including ER81*, ER-α•estrogen* and c-MYC*, all assumed here to be dimerized and active in the phosphorylated (*) state. Other aspects of the model, such as the translation, complexing, and degradation of all mRNAs, proteins, and their complexes, were as described in Appendix A. The final TRN network consisted of 221 genes and 1215 TF/gene interactions, of which 971 were up- and 244 were down-regulating interactions.

A cell state bifurcation diagram constructed for the above epithelial TRN is presented in Fig. 2. This was done to study the effect of the overall level of transcriptional activity as might be imposed by drugs, varying the temperature, changing factors controlling ATP levels (e.g. sugars or O_2), or limiting the available nucleotide pool. Fig. 2(a) shows a complex bifurcation nexus. Here, the root-mean-square (RMS) RNA level (averaged over all genes) is plotted vs. the transcription rate forefactor κ. This type of diagram provides a single measure that elucidates the activity of many genes simultaneously. At a fixed value of κ, Fig. 2(a) suggests there can be as many as nine states. It is likely, however, that roughly half of these states are unstable and several may involve spontaneous temporal oscillations.

To appreciate the biological implications of these many states, the bifurcation diagram must be examined gene-by-gene. The projection of the bifurcation diagram onto the plane of overall rate forefactor and the RNA expression levels transcribed from ESR1 and TP53 are seen in Fig. 2(b). These genes are considered as “key” genes in many cancer related problems. Figs. 2(c) provide the cell state bifurcation diagram as seen for an individual gene (TBP). The graph shows at least one multiple-state feature that appears in different intervals of transcription rate forefactor values. It also displays a curious multiple-state feature in the zone of forefactor values between 0.36 and 0.45 nM/h (the same zone as Fig. 2(a)).

In order to quantify the notion of the complexity of a cell state bifurcation diagram, we introduced several variables. These variables were chosen to elucidate the multiple dimensional character of the structure of the cell state diagrams. According to the model of Appendix A, we investigate bifurcation structure in the N_g+1 dimensional space of the number of genes and the single parameter κ. The notions introduced below are designed to be directly generalizable to a higher number of descriptive variables and bifurcation parameters.

First we introduce complexity measures that are “local,” i.e., that apply at a fixed value of κ. For a given gene we compute the variation of the associated RNA level on the entire range of κ. If this variation is smaller than a prescribed number then that gene is termed invariant and is dropped from further consideration. For a significantly varying gene at a given value of κ, we compute the number of features—i.e. the number of times that the RNA level curve for that gene passes through the fixed κ vertical line at a place that is more than 1% of the total variation of that RNA level over the whole κ range above the intersection just below it. The number of such significant crossings for a given gene, summed over all genes that are not invariant, is the local number of features. Similarly, the number of genes contributing to the complexity is the number of genes for which the maximum minus the minimum RNA level crossing points at a given value κ exceeds 1% of the overall range of values for the given gene on the full range of κ. These measures are plotted in Fig. 3(a,b), and display four bifurcation structures between transcription rate forefactor of 0 and 0.5 nM/h.

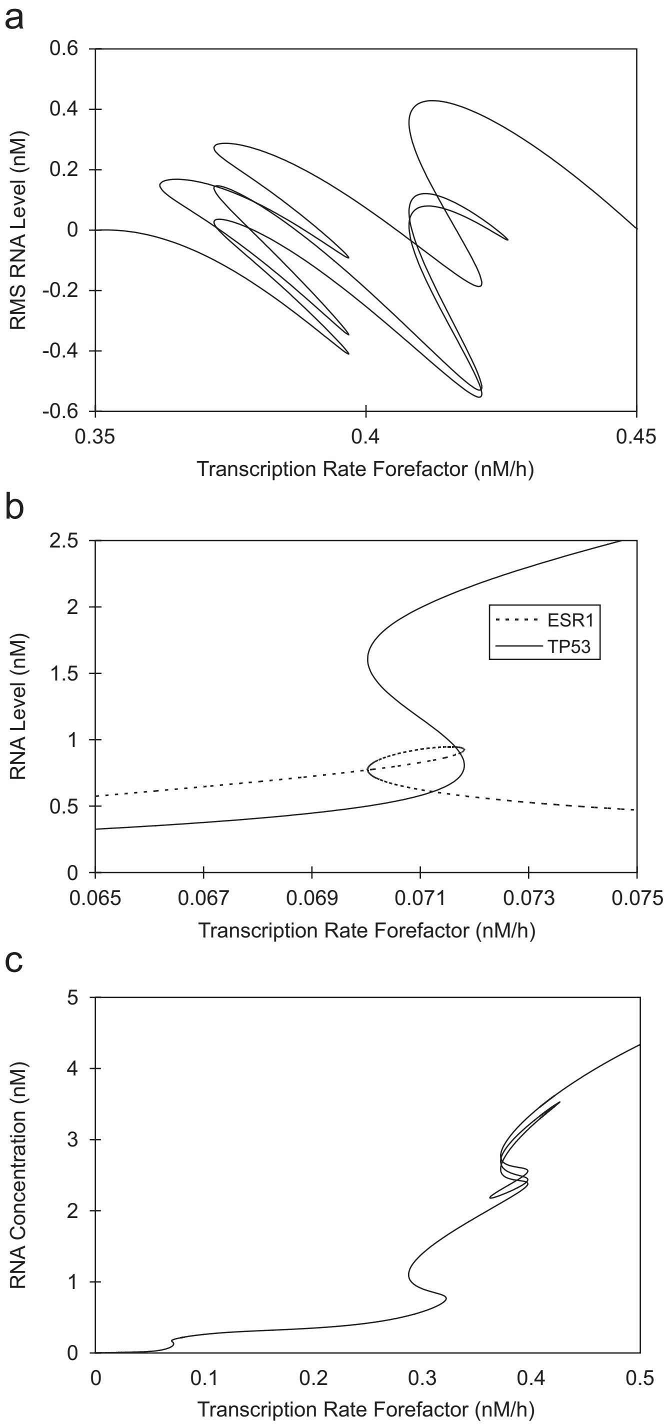

Fig. 2. (a) Feature shows the remarkable complexity that we describe as a bifurcation nexus. The fuller depiction is actually more complex as several branches have bifurcation points from which other branches of steady or oscillatory states emerge. As the RMS RNA is shown, crossing of two branches does not necessarily indicate a branch point—i.e. the fuller depiction is a curve in a roughly 900 dimensional space of the variables (i.e. population levels or binding site occupation, RNA levels, etc.) of the model of Appendix A for this many-gene system. For clarity, we have subtracted a baseline RNA = 76 + 146 [rate factor] from the total RMS RNA. (b) Cell state bifurcation diagram showing two single RNA levels as a function of transcription rate forefactor. The RNAs of interest are TP53 (the solid curve) and ESR1 (the dotted curve). The region of the rate factor (x-axis) is in the region where the total RNA level has the first s-shape. (c) Cell state bifurcation diagram showing RNA level of TBP as a function of transcription rate forefactor. In an area ranging from transcription rate forefactor 0.36–0.45 nM/h, TBP shows a complex feature of cell states.

For local complexity measures, a zone of features can be identified. We also constructed a third measure of complexity that gives information about each of these zones. For a given gene, we compute the integrated area of variation as suggested in Fig. 4. All genes with significant area of variation for a given feature are recorded in Table 1 for the present problem. From this table, it is seen that the bifurcation features (associated cell states) and transitions among them, involve multiple genes simultaneously. Such a collection of genes, when involved in a transition to cancer, might well be termed an onconetwork. When the network of genes and cell states involved was especially complex, we called it an onconexus. The first, third, and fourth zone identified in Table 1 seem to be examples of a bifurcation nexus. Many of the genes involved in the four zones noted in Table 1 are commonly cited as oncogenes or tumor suppressor genes.

Fig. 5 shows the RNA level of an oncogene (JUN red), a tumor suppressor gene (BRCA1 blue) and an auxiliary gene (ATF1 green) as a function of transcription rate forefactor. JUN RNA shows a rapid increase and then a sudden decrease in concentration as the transcriptional rate forefactor increases. The rapid increase in concentration of BRCA1 RNA follows the increase of JUN RNA, and stays almost unchanged except for two small s-shaped features in regions near 0.3 and 0.4 nM/h. The ATF1 RNA shows almost the same behavior as the BRCA1 RNA. From the information in our GeneDat database, we know that genes JUN, BRCA1 and ATF1 form a feedback loop. The gene JUN encodes TF C-JUN/C-JUN which up-regulates gene BRCA1; BRCA1 encodes the TF BRCA1/BRCA1 which up-regulates gene ATF1; gene ATF1 then encodes the TF

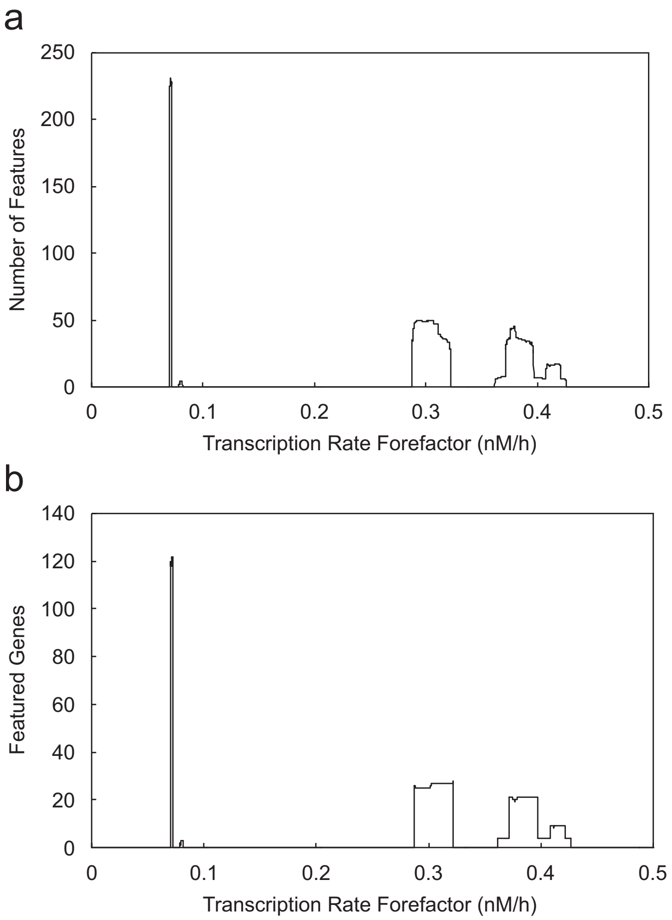

Fig. 3. (a) Number of bifurcation features along the transcription rate forefactor axis showing 4 areas that support multiple cell states. The number of features is computed as the number of distinct RNA concentration levels at each transcription rate forefactor value. The criteria of distinction is to exceed 1% of the overall variation (maximum minus minimum concentration over the full range of the transcription rate forefactor for that gene). (b) Number of genes considered to have distinguishable features at each point along transcription rate forefactor axis showing four areas having multiple states. The criterion of significance for a gene at a given transcription rate forefactor relies on whether there are distinct states (i.e., that the RNA concentration difference is larger than 1% of the variation in RNA level (maximum minus minimum concentration over the full range of transcription rate forefactor for that gene)).

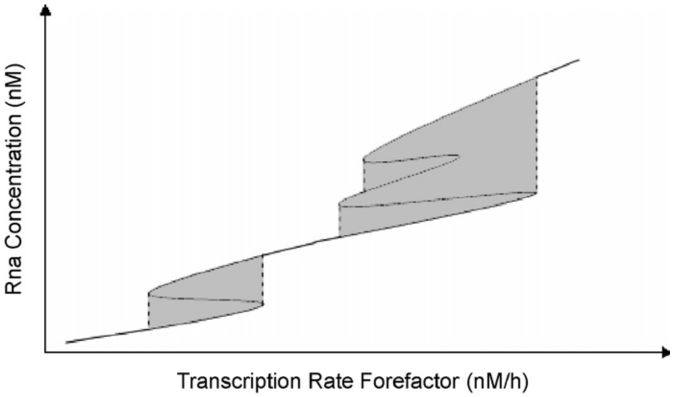

Fig. 4. The "importance factor" of s-shaped states is defined to be the area of the shaded regions shown.

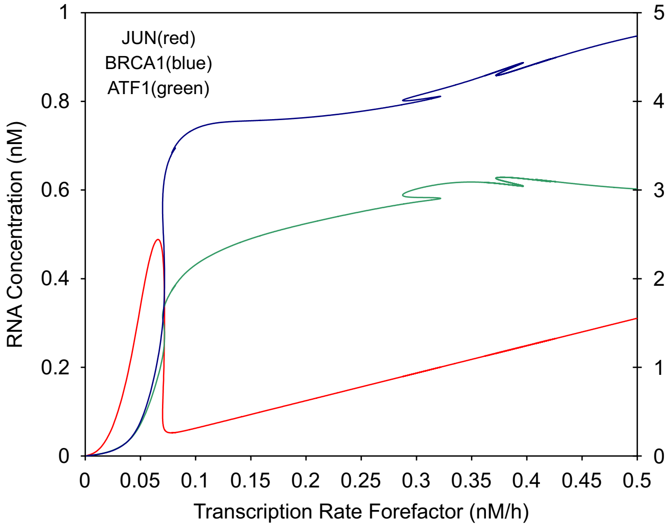

Fig. 5. Cell state bifurcation diagram showing RNA level of an oncogene (JUN red), a tumor suppressor gene (BRCA1 blue) and an auxiliary gene (ATF1 green) as a function of transcription rate forefactor. Curves JUN and ATF1 use the left vertical scale, while the curve BRCA1 uses the right vertical scale. This diagram illustrates the behavior of a coupled oncogene, tumor suppressor gene and auxiliary gene.

ATF1/ATF1 which down-regulates gene JUN. This explains the behavior show in Fig. 5. In addition, this feedback loop could also be considered as a typical motif in an onconetwork. The up-regulation of a tumor suppressor gene by an oncogene might suggest a cell self-protection mechanism, and the indirect down-regulation (through an auxiliary gene) of an oncogene by a tumor suppressor gene indicates the necessity of studying the whole genome-wide TRN instead of focusing on some individual oncogenes.

3. Conclusion and extensions

A supercritical mass of transcriptional regulatory information, augmented with automated TF-based cDNA microarray data analysis, and integrated with bioinformatics modules, appears to hold great promise for achieving a genome-wide understanding of human cell behavior. Through an NDSs discovery approach and cell models automatically generated as noted in Appendices A and B, we have illustrated the feasibility of a workflow taking multiplex bioanalytical data as input and generating predicted cell behaviors as output. In this way, transitions to abnormal states and the structure of the TRN underlying them can be discovered.

Human cell behaviors were shown to emerge as distinct states of the same TRN NDS. This shows that these states can be a consequence of nonlinearity supported by the human TRN. Even for a relatively small human gene subnetwork (i.e. a few hundred genes), we have shown the emergence of many distinct cell states (types), each associated with a distinct cell line and all supported by the same TRN. Although these data are representative, we believe we have demonstrated the usefulness of our integrative, genome-wide approach. We believe that the rich variety of cell types warrants a new term which we propose to be "bifurcation nexus." This term takes on particular significance due to the intense cross-linking of the human TRN. Thus, the many cell states cannot be understood as the juxtaposition of a number of quasi-independent NDSs, each of which supports a few states of a largely isolated sub-network.

Cell transformation has been shown to be a natural consequence of the buildup of otherwise insignificant point mutations and the prevalence of cross-linked feedback in the genome-wide TRN. Thus, even if the structure of the TRN remains unchanged, accumulating point mutations (manifested as small changes in rate and binding constants) can result in an avalanche wherein a cell makes a dramatic genome-wide transition to a qualitatively distinct cell state with the activation of alternative pathways, hence influencing biochemical processes and response to disturbances in the extracellular milieu.

The NDS approach implies that apparently irreversible transitions among cell lines may be prevented or reversed by simultaneously manipulating a number of factors in the extracellular medium. This has potentially dramatic health science implications. For example, a transition to cancer could be arrested or reversed by a change in lifestyle and a transition to drug resistance (e.g. to an anti-estrogen) could be prevented by a change in drug regime or diet. Alternatively, a chemotherapeutic agent could inhibit one transition to cancer, only to promote another due to the complex, often circuitous, network of gene regulatory interactions. The apparent irreversibility of transitions (e.g. Fig. 2(a–c)) could be overcome by the simultaneous application of multiple drugs or by a complex sequence of treatments. Hence drug cocktails, and the sequence and timing with which the drugs are administered, indicate a promising research direction and one which could be greatly enhanced by a genome-wide TRN perspective.

Deriving the clinical implications of the present approach requires a careful examination of the nature of the cell states and transitions between them. The implications of manipulating the activity of apoptotic or cell cycle genes can be derived by examining the cell state bifurcation diagrams, the quantitative information as in Table 1, and the stochastic cell line trajectories (and statistics of cell populations via the latter). The effect of evolutionary procedures could be tested by carrying out Monte Carlo studies for an ensemble of cell line trajectories. In a coming world wherein microarray or other techniques could allow the automated generation of genome-wide, patient-specific TRNs, an automated approach such as that presented here could enable the forecasting of specific transitions to diseased states and allow for the identification of strategies to avoid or forestall these transitions, and to reverse them if they have already been made.

The approach presented is readily extended. The triangular network of Fig. 1(a) (on which the present study was based) could be upgraded to the tetrahedral one of Fig. 1(b). In that case, cDNA microarray data must be supplemented with NMR and proteomics data. With this and the addition of more regulatory mechanisms to our modeling approach, we believe that the methods and workflow demonstrated here will provide practical clinical benefits.

In conclusion, we suggest that the overall automated workflow (microarray data to cancer cell biology) holds great promise for organizing the search for cancer treatments. Basic concepts such as bifurcation nexus and onconetwork (large numbers of genes locked in a strongly coupled network of cancer-related activity) may yield new ways of categorizing cancers. The genome-wide perspective we introduce should help in avoiding adverse side-effects and drug resistance, and in designing multi-target drug cocktails wherein low doses of several drugs can have a much greater effect than a high dosage of a single drug.

Given the automated character of our cell state discovery workflow, it becomes feasible to utilize the large archives of cancer cell microarray data to calibrate and validate our approach. Archives such as NIH's GEO, EBI's EMBL, Oncomine, and SBCR could be utilized and selected subsets of microarray data (e.g. for various epithelial cell lines) investigated to characterize the distinct features of their TRNs. As we believe that the additional insights from this data will enhance the considerable national investment made to date, the type of workflow we have developed should be integrated with these archives so that any new data could be immediately examined, cell state behaviors predicted, and the distinct TRN features of the cell line involved discovered.

Acknowledgments

This project was supported by the United States Air Force (through the Defense Advanced Research Projects Agency) and the United States Department of Energy (Genomics: Genome to Life Program), as well as the College of Arts and Sciences and the Office of the Vice President for Research (Indiana University) as general support of the Center for Cell and Virus Theory, and the IBM Life Sciences Institute of Innovation.

Much of the data on transcriptional regulatory interactions we added to out database by T. Burger, D. Basu and G. Lin.

Appendix A. Model formulation

Considering the general and genome-wide nature of the approach we are attempting to achieve, it is critical to formulate the cell dynamics in a rather general and systematic fashion. For example, the structure of the transcriptional kinetic equations should be the same for all genes; differences between genes are then reflected in the matrix of stoichiometric coefficients introduced and the values of rate and binding constants used.

Gene regulation is described here as a chemical kinetic network involving the attachment/detachment of TFs at sites on each gene. Let P_{ij} be the probability that *j*th site on gene *i* is occupied. Site *ij* is assumed to only be available to a specific subset of TF types (i.e. competitive binding is included and there can be multiple sites for a given TF-type on one gene). In a more general approach, one may introduce the joint probability for each gene such that site–site interaction is accounted for, although this is not done in the present work. If each site is considered to be independent, however, then the kinetic model used here takes the form

$$\frac{\mathrm{d}P_{ij}}{\mathrm{d}t} = k_{ij}^{+} T_{ij}^{*}(1 - P_{ij}) - k_{ij}^{-} P_{ij}, \tag{A.1}$$

where T_{ij}^{*} is the total intra-nuclear concentration of all TFs that bind to site *ij*, and all of which bind to site *ij* in an energetically similar manner. Letting proteins, protein complexes and TFs be labeled via an index *n*. Then

$$T_{ij}^{*} = \sum_{n}^{ij} T_n, \tag{A.2}$$

where T_n is the concentration of the *n*th factor and (*ij*) on the sum indicates a limitation to all factors that can bind and regulate at site *ij*.

Let b_{ij} indicate the nature of the regulation of gene *i* by the TFs binding to site *j*:

$$b_{ij} = \begin{cases} +1, & \text{up-regulation,} \\ -1, & \text{down-regulation,} \\ 0, & \text{no regulation.} \end{cases} \tag{A.3}$$

Introduce a function $\Psi(P, b)$ such that

$$\Psi = \begin{cases} P, & b = +1, \\ 1 - P, & b = -1\,, \\ 1, & b = 0. \end{cases} \tag{A.4}$$

Assuming that a gene is most conducive for transcription if its up-regulating sites are occupied and its down-regulating ones are not, the probability Θ_i, that gene *i* is conducive, is taken to be given by

$$\Theta_i = \prod_{j=1}^{N_{(i)}} \Psi(P_{ij}, b_{ij}), \tag{A.5}$$

where $N_{(i)}$ is the number of regulatory sites on gene *i*. With this, it is assumed that the dynamics of the cellular RNA content R_i for the single RNA type assumed to be associated with gene *i* (i.e. multiple splicing alternates is ignored) is generated by

$$\frac{\mathrm{d}R_i}{\mathrm{d}t} = A_i - \lambda_i R_i, \tag{A.6}$$

$$\frac{1}{A_i} = \frac{1}{k_i^{max}[RP]\{\Theta_i + \zeta_i\}/(1 + \zeta_i)} + \frac{1}{A_i^{poly}} \tag{A.7}$$

for transcription rate A_i and degradation rate coefficient λ_i. The first contribution to $1/A_i$ is due to the rate of RNA polymerase binding to the gene while the second is due to polymerization. Thus, $A_i^{poly} = q_i/L_i$ where q_i is the rate of nucleotide addition during elongation and L_i is the number of nucleotides to be added to make gene *i*-associated RNA. The form of A_i reflects the serial nature of transcription. RNA degradation can take place by several simultaneous processes. Thus we write $\lambda_i = \lambda_i^{(1)} + \lambda_i^{(2)} L_i$. The second term is assumed proportional to L_i, reflecting the possibility that the number of sites on the *i*th RNA at which it can be cut is proportional to its length, while the $\lambda_i^{(1)}$ term accounts for a single or few special sites for initiating degradation which are at endpoints or are relatively inaccessible most of the time due to the RNA conformation. As noted earlier, a cell line is specified via calibration of the k_i^{max}, λ_i, and other parameters. Thus, if there is only one RNA type for each gene under the range of conditions of interest, the above parameters are to be calibrated for these conditions. If the splicing is variable over the range of these conditions, then a multi-channel transcriptional model is required. In (A.7) k_i^{max} is the maximum rate, while ζ_i is a small parameter that allows a minimal rate of transcription even if gene *i* is not optimally conducive, and [*RP*] is the concentration of intra-nuclear RNA polymerase.

Let molecular type *n* be encoded by gene I_n, or arise from a dimerization or other complexing of subunits. For example, estrogen, ERα, ERα•estrogen, (ERα•estrogen)$_2$, etc. are all considered as factors, many of which are related through a network of complexing and other post-translational processes. Let the result of these latter processes have net rate W_n for the *n*th factor. With this, the model

$$\frac{\mathrm{d}T_n}{\mathrm{d}t} = \alpha_n R_{I_n} - \beta_n T_n + W_n(\underline{T}, \underline{c}) - U_n \tag{A.8}$$

is adopted for rate coefficients α_n and β_n, and set $\underline{c}$ of concentrations of other factors (e.g. phosphate, glucose, etc.). When a TF is simply a translated polypeptide, $W_n = 0$; when it arises out of dimerization or other complexing, then $\alpha_n = 0$; when T_n does not regulate gene *i* then $U_n = 0$; this imparts a rather general structure to the model that facilitates implementation of the NDSs analysis. The contribution U_n, which accounts for binding of TFs to regulatory sites on the N_g genes, takes the form

$$U_n = \sum_{i=1}^{N_g} \sum_{j=1}^{N_{(i)}} {}^{(n)} \left\{ k_{ij}^{+} T_{ij}^{*}(1 - P_{ij}) - k_{ij}^{-} P_{ij} \right\}. \tag{A.9}$$

The (*n*) restricts the *j* sum to sites where TF *n* binds.

A number of assumptions have been made in arriving at the present model:

- T_n, as it affects the probability Θ_i, is a concentration and not a thermodynamic activity as might be more accurate due to the concentrated conditions within the nucleus.
- For eukaryotic cells, the factors k_i^{max} and α_n depend on intra-nuclear nucleotide and cytoplasmic amino acid

Table 2
Parameters used in AUTO

	Zone 1		*Zone* 2		*Zone* 3		*Zone* 4	
1	STAT1	44	MYBL1*	20	NFE2L1	807	HIF1A	1693
2	TP73^{+}	27	SMAD3	5	FOXO3A	546	NFE2L1	1609
3	FOXO3A	25	POU2AF1	3	TBP	285	STAT3	311
4	NR3C1	25	NR3C1	3	GATA1	267	GLUR	294
5	BRCA1^{+}	25	POU2F2	2	ATF4	190	TBP	285
6	TP53^{+}	24	CEBPB	1	NR3C1	143	STAT5B	163
7	STAT5B	19	ZNF148	1	USF2	139	STAT5A	147
8	RELA	18	TFEB	1	JUND	132	STAT2	147
9	NFKB1	17	BRCA1^{+}	1	DBP	132	CEBPD	83
10	STAT2	16			TCF4	127	USF2	81
11	STAT5A	16			TBX2	108	JUND	79
12	P63	14			KLF2	105	TCF4	79
13	TNFRSF25	14			JUNB	83	TBX2	73
14	ZNF148	13			ZNF148	82	FOXO3A	70
15	POU2AF1	10			NFKB1	68	KLF2	65
16	FOSL2^{+}	8			FOSL1^{+}	59	JUNB	55
17	RBL2	8			RUNX1	49	ZNF148	42
18	PAX3	7			TFEB	45	NR3C1	30
19	JUN*	6			MYBL1*	33	FOXA3	23
20	ZNFN1A1	6			FOXA3	32	RUNX1	23
21	SP4	6			BRCA1^{+}	25	NFKB1	21
22	RXRA	6			ELF1	23	FOSL1^{+}	18
23	g_+)	6			NFYB	23	BRCA1^{+}	14
24	g_-)	6			FOXA1	21	RXRA	11
25	g_AGE2	6			RXRA	17	MYC*	10

levels, respectively, and they (as well as $[RP]$) may vary in time.

- The description is a lumped, whole-cell model so that T_n and R_i represent numbers of molecules per whole cell volume; the model does not distinguish two populations of TFs (i.e. in the nucleus and in the cytoplasm). Also, the time delays for exchange of TFs and other factors between the nucleus and cytoplasm are ignored.

The above model was implemented for NDSs analysis to discover pathways of cell transformation. To do so, we developed a program that automatically writes the differential equations as FORTRAN-executable expressions. The parameters $k_i^{max}, [RP], \lambda_i, k_{ij}^{+}, k_{ij}^{-}, \alpha_n, \beta_n$, the matrix $\underline{b}$ and vector $\underline{I}$ are user provided. These were obtained in the present study as estimates either from the literature or via our microarray data-based calibration software (see Appendix B and Table 2).

Appendix B. TRN discovery system

The complexity of a mammalian TRN and the limitation and inaccuracies of cDNA microarray data imply that a multi-faceted approach to TRN discovery is required. We have developed a preliminary version of such a system, implemented as a web-based service (sysbio.indiana.edu). Our workflow starts with cDNA microarray data (gene expression profiles) to identify those genes whose expression change during the phenomena of interest. This list of responsive genes is used to initiate a query to our GeneDat database of over 13 000 experimentally verified TF/gene up/down regulatory interactions for mammalian (mostly human) cells. This action also creates a file specifying the gene that encodes each component making up each active TF, as well as hormones or other factors affecting TF activity. In sum, this workflow yields an a priori TRN consisting of the responsive genes, TFs that control them, and the stoichiometry of associated post-translational processes. The scope of the TRN can be extended at the discretion of the user (e.g. adding genes that encode the TFs that regulate the genes that encode the TFs regulating the primary responsive genes). In addition, the TRN assembled contains information on TF complexing and activation. This a priori network serves as the training set for several additional TRN construction modules that add more TF/gene interactions. These modules now include a TF-based method using cDNA microarray data to correct, and extend and calibrate a TRN, as well as GO and promoter analyses to discover new TF/gene interactions. Each method provides a score for every TF/gene interaction it predicts. With this score and a Bayesian method, we compute the ratio of the probability for each score in the training set to that in the random set. Multi-method integration is attained by assigning a sum-log measure from the Bayesian ratios for each TF/gene interaction. A sum-log cutoff is adopted so that only the TF/gene interactions with the highest confidence are accepted. In this way, we have been able to more than double the number of TF/gene interactions over that in the training set, even when the confidence cutoff is taken to be high.

The TRN constructed as above can be further refined by using it as the a priori network for our cDNA microarray analyzers to ensure that the TRN is as consistent with the expression data as possible. For example, to establish a large TRN some of the data assembled from GeneDat may be from abnormal cells, different cell lines, or even nonhuman cells; our microarray-based refinement module is thus critical to screen out spurious data. Finally, our KAGAN cDNA microarray interpreter provides estimates of kinetic and binding constants for TRN processes of Appendix A. Below we briefly review FTF and KAGAN.

FTF microarray-based TRN construction

FTF is our statistical, TF-based module for discovering the structure of a TRN. The input to FTF is an a priori TRN. The output is suggestions for improving the network. FTF is based on the following notions:

- Gene expression data is usually error-prone and thus a consensus method is needed whereby results from a variety of genes are synthesized to derive information on a given gene.
- A method based on TFs has the advantage that microarray noise and error in a user-supplied TRN can be overcome by statistics—i.e. the regulation of many genes can be through a given TF, or a small subset thereof.
- Due to data uncertainty, it is usually not likely (except in rare cases where hundreds of microarrays are available) that there is enough information to obtain both TRN structure and the associated transcription and RNA degradation rate coefficients simultaneously (see KAGAN below, however).
- Network discovery requires many automated trials of possible networks so the algorithm must be extremely efficient.
- Thus, the objective of FTF is to discover TRN structure by taking advantage of the statistical robustness allowed by a TF-based analysis.

The essential FTF equations are as follows. Consider a system with N_g genes. Then

$$T^r_n - T^s_n = \sum_{i=1}^{N_g} H(m^r_i - m^s_i) b_{in} \Psi_{in},$$

where T^r_n is the activity of TF n at condition or time r; m^r_i, the cDNA microarray response for gene i at condition r; b_{in}, the regulatory network matrix ($b_{in} >$ or < 0 for gene i up/down regulated by TF n; b_{in}, the 0 for no regulation); $H(x) = \pm 1$ for $x >$ or < 0, $= 0$ for $x = 0$; Ψ_{in}, the normalized weight—e.g.

$$\Psi_{in} = \frac{2^{L_i}}{2^{L_i} - 1} \frac{1}{M_n}$$

for $L_i =$ number of TFs regulating gene i, and M_n is a normalized factor that is the number of genes TF n regulates.

The advantage of this analysis is that the T^r_n are obtained directly, i.e. no differential equations for them must be solved (as would be the case for a chemical kinetic model, see KAGAN below). The statistical weight Ψ_{in} accounts for the likelihood that a gene controlled by many TFs will not reflect the activity of any one of them. Finally, $\sum_{i=1}^{N_i} \Psi_{in} = 1$, i.e. Ψ_{in} has the character of a normalized probability.

One can compute T^r_n for TF n from one of the genes it regulates by keeping the $T^r_n, (n' \neq n)$ as obtained above. The correlation of this T^r_n and the one constructed as above gives a measure of how well the regulation of gene i by TF n is characterized by b_{in}. Such considerations are the basis of the gene ranking in FTF (and KAGAN—see below). As FTF is fast, many alternative b_{in} can be tested and improvements on the *a priori* TRN network are suggested.

KAGAN microarray-based TRN calibration

KAGAN (Karyote Gene Analyzer) is designed to refine and calibrate a TRN via cDNA microarray data integrated with transcription kinetic modeling via information theory. The basic idea is that the rate of transcription depends on TF activities, thus a chemical kinetic model of RNA level time courses needs TF activity time courses in order to solve the equations. However, RNA levels can be monitored by cDNA microarrays. Using the approach of Sayyed-Ahmed et al. (Sayyed-Ahmad et al., 2003), we introduced an error measure (i.e. observed versus predicted RNA levels) to constrain the probability as a function of model kinetic and binding constants, and a functional of the TF activity time courses. We then find the most probable value of the aforementioned parameters and TF activity time-courses. The resulting equations for the most probable quantities are solved numerically and implemented as the KAGAN module (available through Bio-SPICE and our website sysbio.indiana.edu).

Appendix C. Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2006.12.002.

Appendix D

(See Table 3)

Table 3

Parameters	Value	Data source
K_b maximum transcription rate as limited by RNA polymerase binding	~0.72 (nM/h)	Slutsky and Mirny (2004). http://www.arxiv.org/abs/q-bio.BM/0402005 (preprint web site:condmat).
K_t transcription rate due to reading and elongating	1.5–8.5 (nM/h)	Tomasz Lipniacki and Pawel Paszek, http://mbi.osu.edu/2003
λ mRNA degradation rate	0.029–1.39 (1/h)	www.biology.uc.edu/sophomore/geneticsf01/lec7.pdf
α translation rate	0.15–0.85 (1/h)	Újvári et al. (2001)
β protein degradation rate constant	0.492 (1/h)	Averaged from the literature
γ_+ rate constant of binding for two TFs to TF	2.0 (1/(nM*h))	http://insilico.mit.edu/Overview.pdf
γ_- rate constant for dimerized TF to dissociate	2.0 (1/h)	http://insilico.mit.edu/Overview.pdf
ζ residual transcription rate factor	0.05	None

References

Baselga, J., Norton, L., Albanell, J., Kim, Y.M., Mendelsohn, J., 1998. Recombinant humanized anti-HER2 antibody (Herceptin) enhances the antitumor activity of paclitaxel and doxorubicin against HER2/neu overexpressing human breast cancer xenografts. Cancer Res. 58, 2825–2831.

Bosc, D.G., Janknecht, R., 2002. Regulation of Her2/neu promoter activity by the ETS transcription factor, ER81. J. Cell Biochem. 86, 174–183.

Bosc, D.G., Goueli, B.S., Janknecht, R., 2001. HER2/Neu-mediated activation of the ETS transcription factor ER81 and its target gene MMP-1. Oncogene 20, 6215–6224.

Doedel, E.J., Keller, H.B., Kernevez, J.P., 1991a. Numerical analysis and control of bifurcation problems, Part I. Int. J. Bifurc. Chaos 1, 493–520.

Doedel, E.J., Keller, H.B., Kernevez, J.P., 1991b. Numerical analysis and control of bifurcation problems, Part II. Int. J. Bifurc. Chaos 1, 745–772.

Goel, A., Janknecht, R., 2003. Acetylation-mediated transcriptional activation of the ETS protein ER81 by p300, P/CAF, and HER2/Neu. Mol. Cell Biol. 23, 6243–6254.

Goel, A., Janknecht, R., 2004. Concerted Activation of ETS Protein ER81 by p160 Coactivators, the Acetyltransferase p300 and the Receptor Tyrosine Kinase HER2/Neu. J. Biol. Chem. 279, 14909–14916.

Hahn, H.-S., Ortoleva, P., Ross, J., 1973. Chemical oscillations and multiple steady states due to variable boundary permeability. J. Theor. Biol. 41, 503–521.

Hannsgen, K.B., Tyson, J.J., 1985. Stability of the steady-state size distribution in a model of cell growth and division. J. Math. Biol. 22, 293–301.

Hervagault, J.-F., Ortoleva, P.J., Ross, J., 1991. A plausible model for reversal of neoplastic transformations in plants based on multiple steady states. Proc. Natl Acad Sci. USA 88, 97–10800.

Kauffman, S.A., 1969. Homeostasis and differentiation in random genetic control networks. Nature 224, 177–178.

Lin, C.-Y., Strom, A., Vega, V.B., Kong, S.L., Yeo, A.L., Thomsen, J.S., Chan, W.C., Doray, B., Bangarusamy, D.K., Ramasamy, A., et al., 2004. Estrogen receptor a target gene and response element discovery in breast tumor cells. Genome 5, R66.

Marty, M., Cognetti, F., Maraninchi, D., Snyder, R., Mauriac, L., Tubiana-Hulin, M., Chan, S., Grimes, D., Anton, A., Lluch, A., et al., 2005. Randomized phase II trail of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epitermal growth factor receptor 2-positive metastatic breast cancer administered as first-line treatment: the M77001 study group. J. Clin. Oncol. 23, 4265–4274.

Mochizuki, A., 2005. An analytical study of the number of steady states in gene regulatory networks. J. Theor. Biol. 236, 291–310.

Nicolis, G., Prigogine, I., 1977. Self-Organization in Non-Equilibrium Systems: From Dissipative Structures to Order Through Fluctuations. Wiley Interscience, New York.

Novak, B., Tyson, J.J., 1993. Modeling the cell division cycle: M-phase trigger, oscillations, and size control. J. Theor. Biol. 165, 101–134.

Obeyesekere, M.N., Tecarro, E., Lozano, G., 2004. Model predictions of MDM2 mediated cell regulation. Cell Cycle 3, 655–661.

Ortoleva, P., Ross, J., 1973a. A chemical instability mechanism for asymmetric cell differentiation. Biophys. Chem. 1, 87–96.

Ortoleva, P., Ross, J., 1973b. A theory of asymmetric cell division (differentiation). Develop. Biol. 34, F19–F23.

Ortoleva, P., Brun, Y., Berry, E., Fan, J., Fontus, M., Navid, A., Sayyed-Ahmad, A., Shreif, Z., Stanley, F., Tuncay, K., Weitzke, E., Wu, L., 2003. Karyote: physico-chemical genome, proteome, metabolome cell modeling system, OMICS. J. Integr. Biol. 7, 169–183.

Pegram, M., Hsu, S., Lewis, G., Pietras, R., Beryt, M., Sliwkowski, M., Coombs, D., Baly, D., Kabbinavar, F., Slamon, D., 1999. Inhibitory effects of combinations of HER-2/neu antibody and chemotherapeutic agents used for treatment of human breast cancers. Oncogene 18, 2241–2251.

Piccart-Gebhart, M.J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., Gianni, L., Baselga, J., Bell, R., Jackisch, C., et al., 2005. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N. Engl. J. Med. 353, 1659–1672.

Pietras, R.J., Pegram, M.D., Finn, R.S., Maneval, D.A., Slamon, D.J., 1998. Remission of human breast cancer xenografts on therapy with humanized monoclonal antibody to HER-2 receptor and DNA-reactive drugs. Oncogene 17, 2235–2249.

Rashevsky, N., 1960. Mathematical Biophysics: Physico-Mathematical Foundations of Biology, third ed., vol. 1. Dover, New York.

Sassone-Crosi, P., Cisson, J.C., Verma, I.M., 1988. Transcriptional autoregulation of the proto-oncogene fos. Nature 334, 314–319.

Sayyed-Ahmad, A., Tuncay, K., Ortoleva, P., 2003. Toward automated cell model development through information theory. J. Phys. Chem. A 107, 10554–10565.

Slutsky, M., Mirny, L.A., 2004. Kinetics of protein–DNA interaction: facilitated target location in sequence-dependent potential. Biophys. J. 87, 4021–4035.

Turing, A.M., 1952. The chemical basis of morphogenesis. Trans. R. Soc. London 237, 37–42.

Újvári, A., Aron, R., Eisenhaure, T., Cheng, E., Smicun, Y., Halaban, R., Hebert, D.N., 2001. Translation rate of tyrosinase determines its N-linked glycosylation level. J. Biol. Chem. 276, 5924–5931.

Vogt, P.K., Aoki, M., Bottoli, I., Chang, H., Fu, S., Hecht, A., Iacovoni, J.S., Jiang, B., Kruse, U., 1999. A random walk in oncogene space: the quest for targets. Cell Growth Differentiation 10, 777–784.