Pacific Symposium on Biocomputing 4:5-16 (1999)

GENETIC NETWORK ANALYSIS IN LIGHT OF MASSIVELY PARALLEL BIOLOGICAL DATA ACQUISITION.

ZOLTAN SZALLASI
*Department of Pharmacology, Uniformed Services University of the Health Sciences,
Bethesda, MD, 20814 (zszallas@mxc.usuhs.mil)*

Complementary DNA microarray and high density oligonucleotide arrays opened the opportunity for massively parallel biological data acquisition. Application of these technologies will shift the emphasis in biological research from primary data generation to complex quantitative data analysis. Reverse engineering of time-dependent gene-expression matrices is amongst the first complex tools to be developed. The success of reverse engineering will depend on the quantitative features of the genetic networks and the quality of information we can obtain from biological systems. This paper reviews how the (1) stochastic nature, (2) the effective size, and (3) the compartmentalization of genetic networks as well as (4) the information content of gene expression matrices will influence our ability to perform successful reverse engineering.

Introduction

The study of genetic networks lies in the border area of molecular biology, computer simulations and the study of complex heterogeneous systems. The theoretical and experimental approach to genetic networks developed rather independently, the scientific interaction often being limited to polite curiosity from both sides. There has been an apparent lack of tangible problems requiring the united efforts of experimental biology and theory, since biological studies have produced a relatively low amount of information that was traditionally analyzed by simple decision trees. The limitations of technology lead to a somewhat tautologous strategy of understanding molecular biology. Genes or gene products were classified whether they had a dominant effect on a given endpoint in a certain phenotypic assay. If they did not, they were discarded as irrelevant. If they did, simple decision trees were satisfactory to predict what would happen if the gene or its product were modified, since the gene had a dominant effect in the given biological assay (by virtue of identification). This circular reasoning and the state of experimental tools of molecular biology did not necessitate dealing with combinatorial issues or complex network analysis. Analysis, decision making and prediction did not need to be delegated from "human thinking" to "machine thinking". In cancer research, for example, all experimental tools are biased towards the identification of dominant oncogenes, and the question of non-dominant cooperating oncogenes has been largely ignored. In this latter case, several genes together induce neoplasia, but individually they have no detectable transforming ability. Identification of non-dominant cooperating oncogenes requires a large body of experimental information to start with, and powerful computer tools to analyze complex regulatory interactions. The lack of both of these requirements directed almost all attention towards dominant oncogenes, which in turn proved to be a fruitful research field, since these genes have readily detectable effects. However, as a result, the complex network nature of gene regulatory interactions has been ignored.

It is only recently, that molecular biologists have began to recognize that, for quantitative analysis of gene-expression patterns, living organisms can be viewed as massively parallel computers. Recent technological developments, such as large-scale gene-expression measurements and the human genome project promise an increase of several orders of magnitude in the amount of experimental information. This sudden surge in actual data has renewed the interest in genetic networks from an experimental point of view. For example, it is manageable now for a single scientist to measure within a few months the expression level of all genes of a given organism, such as the approximately 6000 genes of yeast, in a time-dependent manner during the cell cycle[1]. Consequently, biologists will produce less biased and significantly larger experimental databases that cannot possibly be analyzed by simple decision trees. Therefore, experimentalists will face the question of how to make use of large-scale gene expression measurements. Can biological interactions be efficiently predicted based on massively parallel measurements? Can genetic network theory, in cooperation with computer simulations, make any useful, testable prediction about phenotypic changes in a given biological system?

Genetic network analysis is expected to help experimental biology in at least two ways. First, massively parallel mRNA or protein quantitation can produce time-dependent measurements, termed expression matrices, on a significant portion of the members of a genetic network. These expression matrices are the result of the underlying regulatory network. Analytical methods, in particular reverse engineering[2, 3], seek to extract information from time-series measurements in order to identify regulatory interactions in these genetic networks. Then the predicted individual regulatory interactions can be experimentally tested. Forward modeling, on the other hand, is expected to produce expression matrices that accurately predict the time-dependent gene-expression values. All forward modeling starts with an empirically determined database. For example, all molecular biologists studying human cells work on the same directed graph representing the human genes and gene products and their regulatory interactions. If the data generated by them is organized into an efficient database with sufficient understanding of the dynamic regulatory interactions, then, at least in theory, this database could predict time-dependent expression matrices given an initial parameter set. Then the predicted gene-expression patterns can be correlated with the experimentally determined gene-expression matrix and the observed phenotype.

Reverse engineering operates on the following principle: let us consider two consecutive gene-expression states of a gene network. The later state is defined either in a deterministic or a stochastic way by the first gene-expression state and the regulatory interactions. The gene-expression changes occurring during the transition from the first to the second state are consistent with a certain set of regulatory rules. Another pair of consecutive gene-expression states will define another set of possible regulatory interactions. As more and more sufficiently independent, consecutive gene-expression state pairs are examined, the set of possible rules consistent with all of them is narrowed down. In an ideal situation, after examining an appropriate set of samples, a single set of regulatory interactions will be defined that can produce all gene-expression state transitions examined. The set of successive gene-expression patterns examined can be considered as the complete information basis of the analysis. The information requirement for successful reverse engineering depends on the actual genetic network. One of the fundamental questions, which recently became the subject of several pioneering studies[2, 3], is how much information is needed to map the gene-regulatory interactions of a biological system.

At the most recent Pacific Symposium on Biocomputing in 1998, computer simulations were presented by Liang et al.[3] to estimate the number of state transitions in a Boolean network which must be analyzed to determine the regulatory interactions with a certain probability. At the same meeting, John Hertz presented a more general quantitative estimate about the amount of gene-expression measurements necessary to perform reverse engineering on discrete networks[2]. Both of these studies depict a rather favorable picture about the potential use of reverse engineering algorithms to determine regulatory interactions in genetic networks. They essentially state that the number of measurements required to perform meaningful reverse engineering is in the realm of experimental realization. These papers, understandably, considered biological systems as Boolean genetic networks, since these provided the only model system so far that produce insightful and tractable numerical simulations about the overall behavior of large genetic networks[4]. The applicability of the results derived from these discrete networks for real biological systems is one of the key issues in the study of genetic networks. As an attempt to bring theoretical and modeling studies and experimental biology closer to each other, we will approach the question of genetic network analysis from the experimental side. In particular, we will examine the following issues: What is the nature of experimental information that we can obtain from actual genetic networks? What kind of limitations can we expect during analysis due to the nature of experimental biological information? We will present a series of factors inherent in biological systems that will limit the amount of information contained in gene-expression measurements, and which will have a profound effect on the applicability of genetic network analysis. These are: (1) the prevailing nature of the genetic network, (2) the effective size of the network, (3) the compartmentalization of the network, (4) the information content of gene-expression matrices.

In agreement with the relevant literature[5], we will use the following terms and definitions. The qualitative or quantitative expression pattern of genes at a given time point defines a "gene-expression state". A series of successive gene-expression states define a "transition path". The numerical values of time-dependent gene-expression measurements form a gene-expression matrix.

1. The prevailing nature of the genetic network:

For analytical purposes, genetic networks can be viewed either as deterministic or stochastic systems. A deterministic network is a rigid system, where the gene-expression state at a given timepoint and the regulatory interactions between them unambiguously determine the gene-expression state at the next timepoint. In such a network, there is only one path leading from a certain gene-expression state to another, since none of the gene-expression states can have two different successive outcomes (Figure 1). In a stochastic system, on the other hand, a given gene-expression state can generate more than one successive

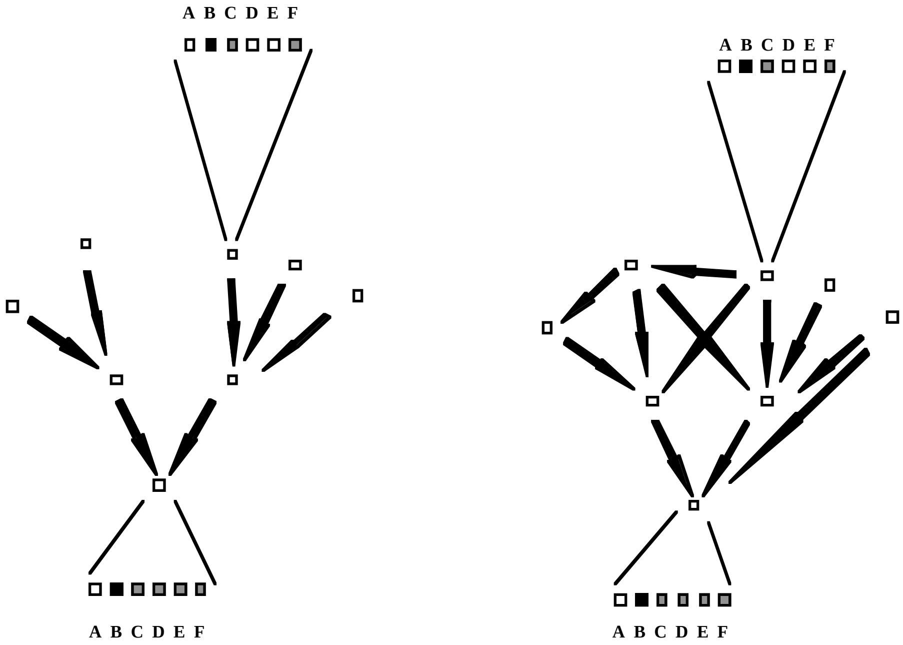

Figure 1. Graphic representation of deterministic and stochastic genetic networks. Every square-shaped node represents an entire gene-expression state, defined by the expression level of thousands of genes. Here we show only six (A to F) genes, displaying levels from no expression (empty squares) to maximum level of expression (full squares). Connective directed "edges" represent state transition paths showing the direction of the transition. Note, that in the deterministic network of figure A there is only one state transition path exiting each state, whereas more then one path can lead into the same gene-expression state. In the stochastic network of figure B, a multitude of transition paths exit each node, and there are several paths entering each gene-expression state as well.

gene-expression states, and therefore, different cells of the same population may follow a different gene-expression path from one state of gene-expression to another. The fact that in reality genetic networks are stochastic, is supported both by theoretical considerations and experimental results. The number of transcription factors in a cell nucleus is often low, about a couple of hundred; the environment in which the gene regulatory interactions occur is far from free solution; and the reaction kinetics is relatively slow. As a result of all these factors taken together, stochastic mechanisms describe the kinetics of gene regulation more accurately than a deterministic description, such as a set of continuous differential equations[6]. In addition, there is a growing number of experimental results both in prokaryotes (reviewed in reference 7) and eukaryotes, such as hematopoiesis in mammals[8] that could be best explained and modeled by stochastic mechanisms. In general, stochasticity allows significant variations in the sequence of activation and inactivation of genes. In extreme cases, this can result in two cells undergoing the same phenotypic change (e.g. proceeding from a certain point in the cell cycle to a later one) but having the sequence of activation for two genes reversed (Figure 2). Conceptually this would not be a problem if we were able to analyze the gene-expression matrixes derived from individual cells, although the computational cost of analyzing a large number of alternative time-series data, considering the stochastic generation of gene-expression matrices is probably high. More importantly, time-series measurements will always be obtained as population-averaged data. Current technology for massively parallel gene-expression measurements requires starting materials derived from a large number, often up to tens of thousands or millions, of cells. Even if it were possible to extract all quantitative parameters from a single cell at a given time point (e.g. using highly sensitive, PCR-based technology) the measurement itself kills or profoundly alters the organism, and we have no idea which transitions the organism would have proceeded through at later time points. In extreme cases, like the one in Figure 2, the population-averaged measurements will mask the real regulatory interactions. In reality, gene X and gene Y might be in a regulatory interaction with each other and never be activated at the same time. The time-averaged measurements, however, would suggest that they are activated at the same time. Stochastic simulations of large genetic networks[7,9] are needed to assess how often we can expect the masking of regulatory information by the above-described mechanism.

Stochasticity can be one of the main reasons for the lack of sharp switch-on and -off kinetics of gene-expression[7], as often observed in experimental systems. This will put an empirical limit on the number of informative timepoints obtained in time-series measurements. In the yeast cell cycle, for example, it takes about 10-20 min from the first traces of gene activation to reach the maximum level of expression. In the human cell cycle, this time interval might increase up to two hours. The error of quantitation for individual genes by traditional methods such as Northern blot hybridization or quantitative PCR is about 15-20%, whereas the massively parallel technologies, such as cDNA microarray or oligonucleotide chips, currently suffer from a somewhat higher (30-50%) measurement error. A rational experimental design will sample time-dependent gene-expression according to a time-series in which each pair of consecutive measurements is expected to produce at least as large expression level difference as the error of the measurement. Therefore, the average speed of gene activation and inactivation in combination with the error of the quantitation method applied will define an optimal time scale of the experiment. More frequent sampling is unlikely to yield additional information and less frequent sampling will likely lead to missing regulatory interactions.

This would yield 5-10 min sampling time intervals for cDNA microarray based measurements in the yeast cell cycle,. The 10 minute time-intervals chosen by Ron Davis's group to cover the cell cycle of S. Cerevisiae followed closely the gene-expression dynamics[1]. Similarly, for human cells, massively parallel measurements could yield a maximum of 50-80 useful timepoints. A similar maximum for the number of informative timepoints can be established for other time-dependent measurements, such as the effects of the activation of a given transcription factor,

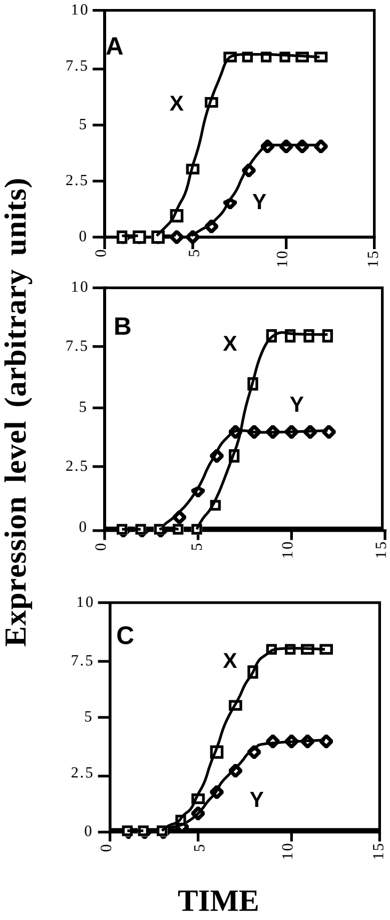

Figure 2. Time dependent expression changes of genes X and Y in two individual stochastic genetic networks (figures A and B) and the experimentally measured, population-averaged values (figure C). As described in the text, stochastic networks can get from a given gene-expression state to another one via several alternative paths. In the genetic network represented in figure A, gene X is activated before gene Y. Another cell (figure B) might follow a gene-expression path in which the sequence of the activation of these genes is reversed, even if gene X and Y are in regulatory interaction. Experimentally, we can measure only the population-averaged values. Based on the expression kinetics displayed by the genetic networks in figure A and figure B, both genes will show the same time dependence of gene-expression changes, although the maximum level of expression remains different.

as well. We will show below, in point 4, how this will effect the amount of maximum information contained in gene expression matrices.

Stochastic genetic network modeling, currently under development by several groups[7,9], could provide an insight into the severity of quantitation problems to be expected during population-averaged time-series measurements.

2. The effective size of the network:

In modeling, genetic networks are often treated as deterministic networks of the genes contained in the given cell. It is assumed that the expression-state of every gene is determined unequivocally by the expression-states of its input genes. It is well known, however, that the regulatory interactions between genes are not deterministic at the mRNA level. A whole series of regulatory events exist between the activation of a certain gene and the effect of the same gene on a downstream regulated gene, and these regulatory events often receive multiple conditional inputs from an array of other elements in the network. Let us review a demonstrative example for the several steps involved from the production of mRNA of a transcription factor until the production of mRNA of a downstream-regulated gene. The product of the gene “c-jun” forms the AP-1 transcription factor complex, either with another identical c-jun molecule as a homodimer, or with one of several other proteins as a heterodimer [10]. We start from the state when the mRNA of c-jun is already produced (for the transcriptional regulation of this gene, see[10]). In addition to the transcriptional regulation, the level of mRNA of this gene can be regulated by the stabilization or destabilization of mRNA[11, 12] (Figure 3). The level of protein production will be proportional to the net amount of c-jun mRNA and not to the transcriptional activation of this gene by itself. mRNA is the first regulated derivative of the c-jun gene. All proteins are produced in the cytoplasm in a non-modified form, and the jun protein has to be first translocated to the nucleus to exert its transcriptional activity. The nonphosphorylated cytoplasmic and nonphosphorylated nuclear jun protein can be considered as two further derivatives of the c-jun gene, since there is evidence for the independent regulation of localization[13]. The activity of the jun protein is significantly enhanced by phosphorylation at serine 73 and serine 63. The nuclear phosphorylated form of c-jun can be considered as an additional derivative, since both the function and the regulation by stabilization differs for the phosphorylated and nonphosphorylated form[10]. For example, the non-phosphorylated form can effectively inhibit the glucocorticoid receptor activity, but it is much less effective in binding to and activating the TPA-responsive element[10]. As shown on Figure 3, c-jun has at least four independently regulated derivatives: its mRNA, the nonphosphorylated cytoplasmic, nonphosphorylated nuclear and the phosphorylated nuclear forms. Biochemistry is constantly increasing our knowledge about the localization of proteins in different subcellular domains and about posttranslational modifications, often at multiple and independently regulated sites of the same proteins. Without doubt, this will further increase the number of independently regulated derivatives of each gene. In addition to the “specific regulatory inputs” of proteins the activity of biological molecules is often regulated by the concentration or concentration

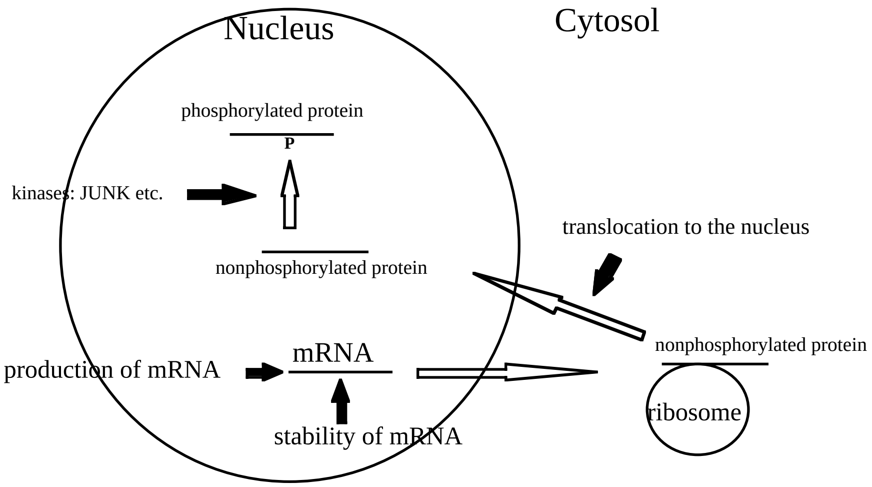

Figure 3. Independently regulated derivatives of the c-jun gene. For details see text.

gradient of small molecules or ions, such as Ca^{++}, as well. Because of these abundant regulatory factors, a genetic network based only on gene expression patterns is non-deterministic. We have recently suggested[14] the introduction of "biological parameters" as the regulatory input unit, that incorporates all the above described gene-derivatives, small molecules, etc. Models in experimental biology have obviously been built on biological parameters instead of genes[14]. For a given gene, mRNA, inactive and active protein, localization of proteins, cofactor dependence, etc. must all be incorporated into the model.

Currently, the biochemical description of a single gene and its derivatives involves about 5 to 10 distinct biological parameters on average. This would result in a deterministic genetic network with about 10 times as many members than the total number of genes.

3. The compartmentalization of the genetic network:

The level of compartmentalization of genetic networks will have a profound effect on the analysis of gene-expression matrixes. A high level of compartmentalization essentially means that the number of regulatory interactions between subgroups of members of the genetic network is low, and therefore the number of regulatory interactions to be tested by reverse engineering algorithms could be significantly reduced. Whereas the effective size of the network will be determined by an ongoing, concerted, and relatively slow effort by a significant portion of biochemistry, the level of compartmentalization could be assessed with a more limited effort. An increasing number of transcription factors, such as several members of the nuclear receptor superfamily, can be efficiently, selectively and quickly activated and inhibited[15]. In fact, small molecules that act as selective transcription factor modulators is one of the main research directions of the

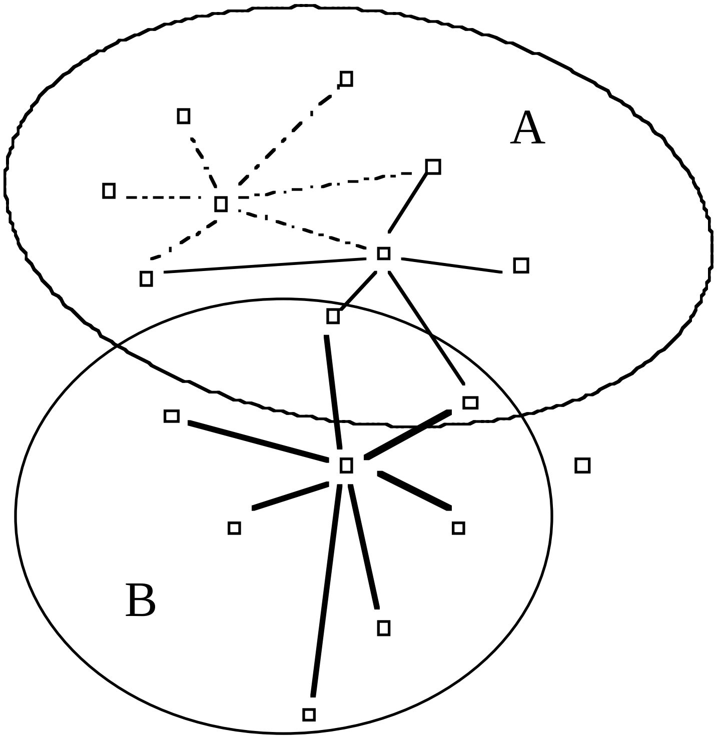

Figure 4. Experimental determination of compartmentalization of a genetic network. The figure shows a hypothetical set of three experiments. Directly regulated genes were identified from gene-expression matrices after the activation of three different transcription factors. Each square represents a gene, and each connecting edge represents a direct regulatory input. The regulatory interactions identified from the different experiments are distinguished by different line types. The gene in the middle of the radial pattern of regulatory interaction is the modulated transcription factor.

pharmaceutical industry, and we could expect a steady increase in the number of these agents. Direct modulation of transcription factors is expected to change the expression level of a significant portion of all of its directly regulated genes. Of course, there will be several directly regulated genes without any change in their expression despite the direct modulation of the transcription factor, because, e.g. an IF NOT function is involved in the regulation and the inhibitor is present, or an AND function is involved and the other input is not present etc. Nevertheless, analyzing the early and late time-points of the gene expression-matrix induced by the direct modulation of a given transcription factor will give an informative quantitative estimate about the number of directly regulated genes and about the number of secondary, tertiary etc. regulated genes, respectively. Let us consider only the directly regulated genes. Figure 4 shows the results of a set of hypothetical experiments. Increasing the number of direct regulatory interactions entered into the database will gradually outline gene groups that have relatively few regulatory connections between each other. In this case, gene groups A and B have relatively few regulatory interactions, although because of the low number of experimental data, several other possible gene compartments could be outlined as well. In case of significant compartmentalization, reverse engineering can take advantage of the low probability of regulatory interactions between different gene subgroups. Therefore, the computational time will be significantly reduced

4. The information content of gene expression matrices:

DNA chip technologies [16, 17], proteomics, etc., are expected to produce a wealth of information about quantitative changes in biological systems. This seemingly vast database might provide surprisingly little information for quantitative analysis. Let us consider for example the eukaryotic cell cycle. The complete time-dependent gene expression matrix has already been produced for the cell cycle of the yeast, S. Cerevisae[1]. Out of roughly 6,000 genes, 416 oscillated. The majority of these genes peaked only once, and 33 peaked twice. Similarly, it can be generalized from the literature for other organisms, that cycling genes tend to peak only once during a complete cell cycle.

As we have seen above, it is not informative to take gene-expression measurements more often than every 5 minutes for yeast and 15-30 minutes for mammalian cells. Therefore, the maximum number of time points of a gene-expression matrix is $T_{min}/5$ for yeast, and between $T_{min}/15$ and $T_{min}/30$ for mammalian cells, where T_{min} is the length of the biological process examined, expressed in minutes. For the yeast cell cycle, this is typically around 90 min and for mammalian cells usually between 900 to 1800 minutes (15 and 30 hours). The gene-expression matrix for yeast can provide in the order of $N \times (T_{min}/5)$ bits of information. The majority of the genes, however, do not change their expression, and the ones that do, change their expression only twice. Therefore, the identity of the cycling genes, and knowing when they are turned on and off, carries all the information about the cell cycle gene expression matrix. Now let us consider the 422 cycling genes versus the total number of genes in yeast and the maximum number of quantitation time-points, which is about 20. Simple calculations show that the actual amount of information obtainable from the yeast gene expression matrix is about 1 to 2 orders of magnitude lower than expected by the size of the gene expression matrix alone. A similar rate of information can be established from the time dependent gene-expression matrix for S. Cerevisiae during the metabolic shift from fermentation to respiration [18]. For higher eukaryotes, we do not have similar comprehensive measurements yet. Nevertheless, a large body of data suggests that the pattern of infrequent oscillation of genes and the ratio of oscillating versus fixed expression genes will be similar in higher organisms[14].

Discussion

In his paper, John Hertz estimated that, for successful reverse engineering, the number gene expression states to be measured, P will be in the order of

(1) P=K log(N/K)

where N is the size of the entire network (e.g. the number of genes) and K is the average number of regulatory interactions per gene[2] . We pointed out, that a deterministic gene-expression network will have to deal with about 5 to 10-fold more variables than the number of genes in the system. According to equation (1), the one order of magnitude increase in the size of the network will at most double the number of experiments required. The information-theoretical perspective of Hertz's calculation also lets us estimate the effect of the reduced level of information contained in gene-expression matrices described in point 4 [19]. According to Hertz, if the expression matrix has only aPN bits of entropy instead of the full PN, then the necessary number of experiments is increased by 1/a. (For details on the calculation, see[2].) We estimated about 1 to 2 orders of magnitude less information in gene-expression matrixes than expected by their size. This will send a more alarming message to experimentalists. A 10-fold increase in cost may be troublesome and a 100-fold increase is often prohibitive in experimental biology. Assuming that the cost per information unit does not decrease significantly with the increasing number of experiments, the information content of gene-expression matrices might impose financial limitations on reverse engineering.

The level of compartmentalization on the other hand might improve our chances for efficient reverse engineering, although the quantitative data are still missing to make estimates about this issue.

Finally, the nature of the genetic network seems to be the most uncertain point in the quantitative analysis. The level of masking useful regulatory information due to stochasticity is unknown. This issue can be hardly assessed by experimental tools because of the very nature of the problem. Modeling large-scale stochastic genetic networks seems to be the most informative approach that could answer several questions. Does the stability, low level of oscillation, redundancy and other overall features of biological systems restrict the state transition paths in a way that the sequence of gene activation will remain the same with a low error?

In summary: During the next few years theory, modeling and experimental data generation is expected to provide reliable estimates for the problems discussed above. At this moment, the question is open: reverse engineering might prove to be a powerful tool to decipher gene-regulatory interactions, but if the above discussed numerical features turn out to be unfavorable, then the lack of information will highly impede the application of this method. In this latter case, genetic network analysis will be restricted to the well-known simple decision making process, which in this case is: If a gene was activated within a certain time frame after another gene, then they might be in regulatory relationship, that could be tested experimentally.

The interconnected nature of the problems presented in this paper will hopefully generate a closer collaboration between the representatives of theoretical and experimental biology.

Acknowledgments:

The opinions and assertions contained herein are the private opinions of the author and are not to be construed as official or reflecting the views of the Uniformed Services University of the Health Sciences or the U.S. Department of Defense.

The author would like to thank Dr. John Hertz (Nordita) and Dr. Shoudan Liang (NASA, Ames Research Center) for several informative conversations and e-mail exchanges. The author also thanks Dr. Gregory T. Klus for critical reading of the manuscript.

References

1. Cho, R.J, Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R. W. *Molecular Cell*, **2**:65-73, (1998)
2. Hertz, J. in *Pacific Symposium on Biocomputing* (Maui, Hawaii, 1998). The paper is available at http://www.nordita.dk/~hertz/projects.html
3. Liang, S., Fuhrman, S. & Somogyi, R. *Pac. Symp. Biocomput.* **1998**, 18-29 (1998).
4. Kauffman, S. *The origins of order* (Oxford University Press, Oxford, 1993).
5. Somogyi, R. and Sniegoski, C.A. *Complexity*, **1**:45-63 (1996).
6. Gillespie, D.T. *J. Phys. Chem.* **81**, 2340-2361 (1977).
7. McAdams, H.H. & Arkin, A. *Proc. Natl. Acad. Sci. USA* **94**, 814-819 (1997).
8. Abkowitz, J.L., Catlin, S.N. & Guttorp, P. *Nat. Med.* **2**, 190-197 (1996)
9. Gibson, M. & Bruck, J. (manuscript in preparation)
10. Karin, M., Liu, Z. & Zandi, E. *Curr. Opin. Cell Biol.* **9**, 240-246 (1997).
11. Ausserer, W.A., Bourrat-Floeck, B., Green, C.J., Laderoute, K.R. & Sutherland, R.M. *Mol. Cell Biol.* **14**, 5032-5042 (1994).
12. Trejo, J. & Brown, J.H. *J. Biol. Chem.* **266**, 7876-7882 (1991).
13. Haase, M., *et al. Virchows Arch.* **431**, 441-448 (1997).
14. Szallasi, Z. & Liang, S. *Pac. Symp. Biocomput.* **1998**, 66-76 (1998).
15. Mangelsdorf, D.J., *et al. Cell* **83**, 835-840 (1995).
16. Chee, M., *et al. Science* **274**, 610-614 (1996).
17. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. *Science* **270**, 467-470 (1995).
18. DeRisi, J.L., Vishwanath, R.I. & Brown, P.O. *Science* **278**, 680-686 (1997).
19. Hertz, J. (1998), Personal communication.