Analysis of Gene Expression Data **Spring Semester, 2004**

Lecture 10: May 27, 2004

Lecturer: Ron Shamir *Scribe: Omer Czerniak and Alon Shalita*[1]

10.1 Genetic Networks

10.1.1 Preface

An ultimate goal of a molecular biologist is to use genetic data to reveal fundamental cellular processes, and their impact on complex organisms. In order to achieve this goal one has to study how complex systems of several genes and proteins function and interact.

10.1.2 Genetic Networks

Definition A *genetic network* is a set of molecular components such as genes, proteins and other molecules, and interactions between them that collectively carry out some cellular function.

Genetic Networks describe functional pathways in a given cell or tissue, representing processes such as metabolism, gene regulation, transport and signal transduction. Let us examine several examples:

1. **Expression of the Gene proB**

   Figure 10.1 depicts the gene's expression and its role in catalyzing a specific chemical reaction in the cell. The proB gene is being expressed into the gamma-glutamyl-kinase protein, which catalyzes a reaction involving glutamate and ATP, that produces gamma-glutamyl-phosphate and ADP compounds.

2. **A Simple Metabolic Pathway - Proline Biosynthesis**

   The next example is part of a simple metabolic pathway, involving a chain of generated proteins, which is shown on Figure 10.2. One of the final products of the chain, proline, inhibits the initial reaction that started the whole process. This "feedback inhibition" pattern is highly typical to genetic networks, and serves to regulate the process execution rate.

[1] Based in part on a scribe by Meital Levy and Giora Unger 2002, Koby Lindzen and Tamir Tuller 2002

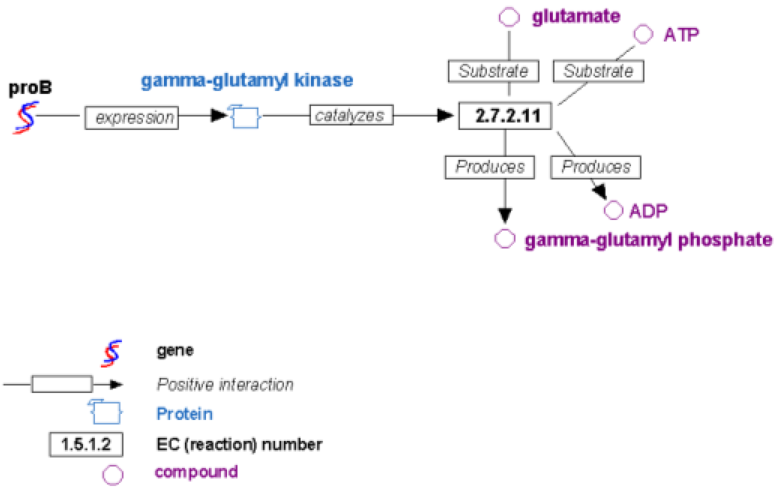


Figure 10.1: An example of the role of gene expression in catalyzing chemical reactions.

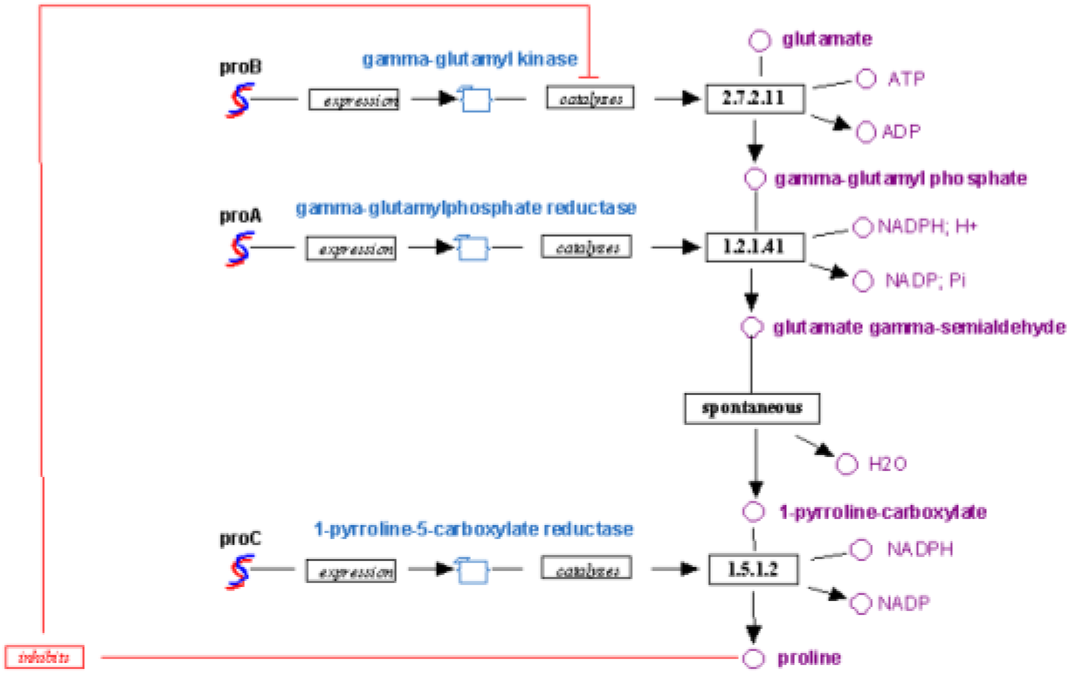


Figure 10.2: An example of a metabolic pathway: Proline biosynthesis.

3. **Methionine Biosynthesis in E-coli.**

   The following two figures show a more complex genetic network, describing Methionine biosynthesis in E-coli. The second figure is a schematic representation of the pathway, with most nodes omitted, but it can give a better idea of the overall topology.

4. **Signal Transduction Network**

   This example, depicted in Figure 10.5, is that of signal transduction - a complex cellular process initiated by a signaling protein, arriving from outside of a cell. This process eventually affects gene expression in both the cytoplasm and inside the nucleus.

10.1.3 Experimental Startegies

Using a known structure of such networks it is sometimes possible to describe the behavior of cellular processes, reveal their function and determine the role of specific genes and proteins in them. That is why one of the most important and challenging problems today in molecular biology is that of *functional analysis* - discovering and modelling Genetic Network from experimental data.

Biological Tools

There are two central approaches in addressing this problem: The first approach tries to find out the relation between two specific genes. An example of this approach is the usage of 2-hybrid systems [2]. The second approach takes ”snapshots” of the expression levels of many genes in different conditions, and according to that, tries to describe the network of relations between these genes. An example of this approach is the usage of DNA microarray, commonly used to monitor gene expression at the level of mRNA. The main contribution of this technology is that numerous genes can be monitored simultaneously, making it possible to perform a global expression analysis of the entire cell. In this scribe we will cover techniques related to the second approach.

Additional information about a genetic network may be gleaned experimentally by applying a directed perturbation to the network, and observing expression levels of every gene in the network, in the presence of the perturbation. Perturbations may be *genetic*, in which the expression levels of one or more genes are fixed by *knockout* (removal of the gene) or *overexpression* (higher than usual level of gene expression), or *environmental*, in which one or more non-genetic factors are altered, such as a change in environment, nutrition, or temperature. Such biological experiments are very costly and very few such perturbations may be performed at one time. Thus, reducing the number and cost of experiments is crucial.

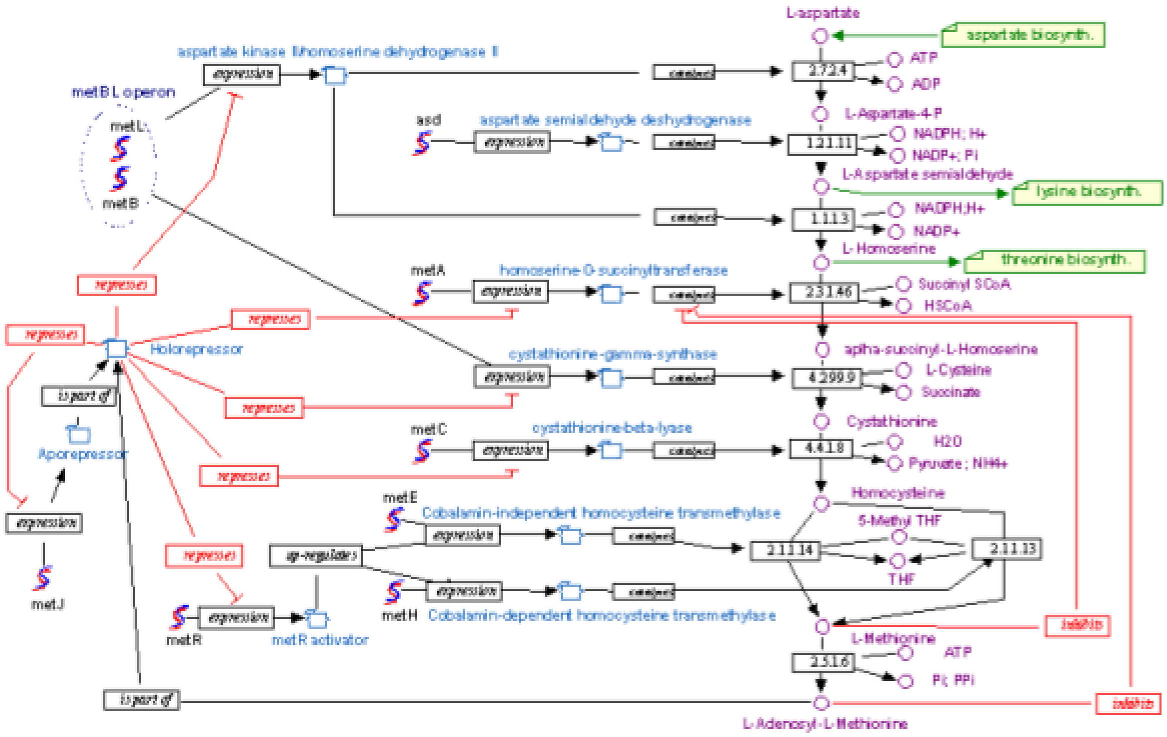


Figure 10.3: Methionine biosynthesis network in E-coli.

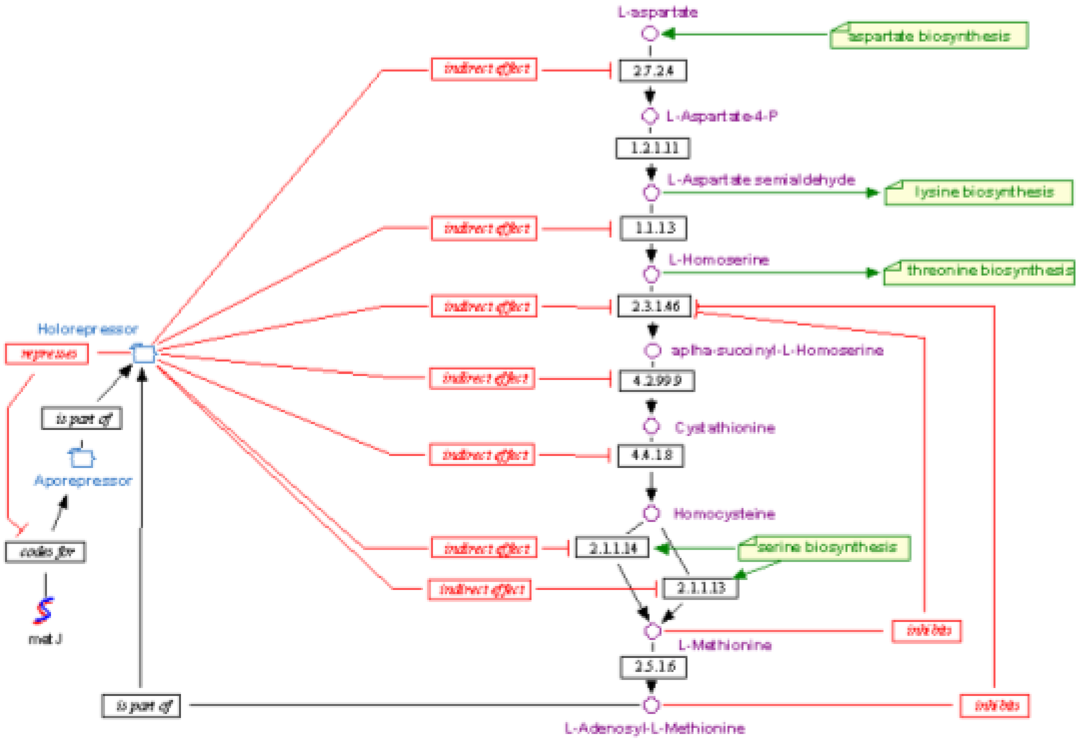


Figure 10.4: Schematic representation of the biosynthesis pathway presented in Figure 10.3.

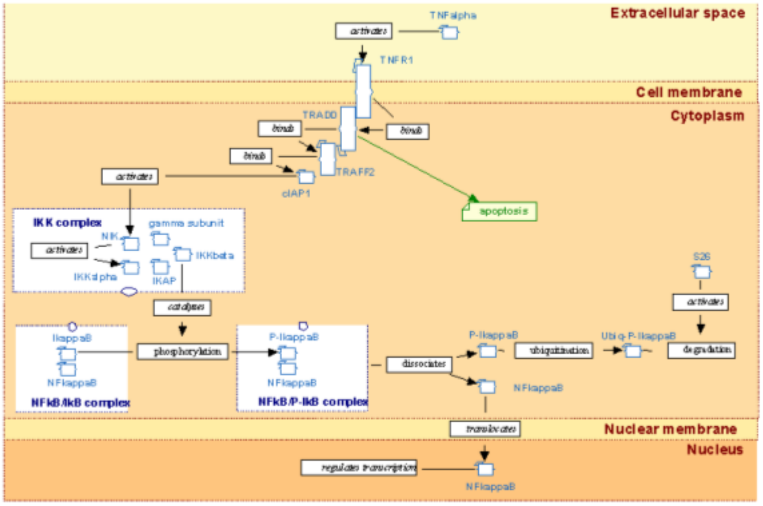


Figure 10.5: A genetic network that performs signal transduction from outside the cell into the nucleus.

The methods presented above supply biological data in terms of expression levels of many genes at different time points and under various conditions. The functional analysis of the data can be defined as a computational problem, aiming to infer some plausible model of the network from the observations, while keeping the number or cost of biological experiments at a minimum. The model should describe how the expression level of each gene in the network depends on external stimuli and expression levels of other genes. Additional goals include construction of a knowledge-base of gene regulatory networks, and verification of pathways or genetic network hypotheses.

10.1.4 Genetic Network Models

In the process of modeling a genetic network, one tries to find out which components are involved in the network and the interactions between them. Several models have been proposed in the literature to capture the notion of genetic networks and allow mathematical solutions of the computational problem of modeling biological processes:

- **Linear Model:**
  This model, proposed by D'haeseleer et al [5, 4], assumes that the expression level of a node in a network depends on a linear combination of the expression levels of its neighbours.

- **Boolean Model:**
  Proposed by Kauffman[7]. It assumes only two distinct levels of expression - 0 and 1.

According to this model, the value of a node at time $t + 1$ is a boolean function of the values of its neighbours at time t.

- **Bayesian Model:**
  Proposed by Friedman et. al [6]. It attempts to model the behavior of the genetic network as a joint distribution of different elements.

In this section we shall concentrate on the Boolean model. The Bayesian Model will be discussed in detail in the next Lecture.

10.1.5 Boolean Network Model

According to the boolean model, a network is represented by a directed graph $G = (V, F)$, where:

- V represents nodes (elements) of the network.
- F is a set of boolean functions (see below), that defines a topology of edges between the nodes.

A node may represent either a gene or a biological stimulus, where a stimulus is any relevant physical or chemical factor which influences the network and is itself not a gene or a gene product. Each node is associated with a steady-state expression level x_v, representing the amount of gene product (in the case of a gene) or the amount of stimulus present in the cell. This level is approximated as high or low and is represented by the binary value 1 or 0, respectively.

Network behavior over time is modeled as a sequence of discrete synchronous steps. The set $F = \{f_v | v \in V\}$ of boolean functions assigned to the nodes defines the value of a node in the next step, depending on values of other nodes, which influence it. The functions f_v are uniquely defined using truth tables. An edge directed from one node to another represents the influence of the first gene or stimulus on that of the second. Thus, the expression level of a node v is a boolean function f_v of the levels of the nodes in the network which connect (have a directed edge) to v.

Definition A *trajectory* is a sequence of consecutive states of the network. It can be viewed as a list of N-dimensional vectors (N being the number of nodes in the network), each representing a state.

10.1.6 A Complementary Approach

We can view an organism as a very large genetic network. If we knew all the interactions of such a network, we could perfectly understand every single detail in the organism. That is, we could understand which genes, proteins and other molecules are involved in every biological process, how exactly the process takes place, etc.

This might be the ultimate goal of biological science, but obviously we are light years away from it. We therefore make a simplifying assumption. We model the organism as many distinct genetic networks, which loosely interact among themselves.

Indeed, this is a heavy assumptions, but it is necessary in order for genetic networks to be useful in modeling biological processes.

Instead of looking at a specific network, we look at general properties of ”network of the kind” (eg. networks where each components has exactly 2 related components). Given such a group of genetic networks, we can explore their properties (global structural features, types of possible dynamic behaviors etc.). The search for generic properties may also provide hints for the analysis of specific circuits (like which features to expect, what questions to ask, etc.).

Definition An *ensemble of genetic networks* is composed of similar networks that share some features. The non constrained features vary at random between networks in the ensemble.

Properties of an Ensemble of Networks:

- Every network consists of N nodes (genes).
- Each gene is influenced directly by exactly k other input genes.
- For each node, the k input genes are chosen at random.
- For each node, its boolean function is chosen at random from the 2^{2^k} possible functions (the table of the input has size of 2^k states, and for each state the function can return 0 or 1).

10.1.7 Simplified Description

Following are a few assumptions taken in order to simplify the model:

- The activation of genes depends on proteins and chemicals.
- The synthesis of proteins participating in a regulatory process is very fast compared to the regulatory process itself.
- Regulatory proteins decay much faster than the duration of the regulatory processes.
- The concentrations of the regulatory chemicals are constant.

As a result of those assumptions, we can express the activation level (mRNA level or protein level) in time $t + \delta t$ as a function of the activation at time t. We will later use $\delta t = 1$. This means that loss of memory occurs within δt time, that is, knowledge of steps before time T is not needed.

10.1.8 Kauffman's Model

Kauffman's Model [7, 8] uses boolean gene levels, 1 for active and 0 for inactive. It also assumes that time $t + 1$ is determined by a boolean function of the levels of a fixed set of input genes at time t. This means it can use only 1-step memory. All updates are executed in a deterministic way and are synchronized. External chemicals are not explicitly taken into account. The module assumes we have N nodes. We choose random topology between the nodes, than we choose random functions betweens the genes that effect a gene (”regulators”) and the gene itself (the ”regulatee”), and than we choose random initial values for the nodes at time 0. See 10.6.

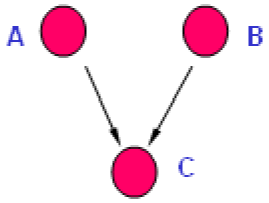


Figure 10.6: $C(t + 1)$ the reulatee depends upon the regulators $A(t)$ and $B(t)$.

Kauffman's Model is dynamic:

- At time 0, a level is given to every gene.
- At each time step $t = 1, 2...$ every gene has a level $x_i(t)$, which is determined according to the boolean functions.
- The global state of the system is $X = [x_1, x_2, ...x_n]$ and we say that $X(t)$ alone determines $X(t + 1)$. As time passes, the system moves from state $X(t)$ to $X(t + 1)$, $X(t + 2)$ and so on, following a trajectory.

The states can be thought of as corners in the unit hypercube and a step from one global state to another can be thought of as shifting from one corner to another. Note, that a legal move does not have to be between two adjacent corners, since adjacent corners differ only by one bit. See 10.7 a 3-dimensional cube.

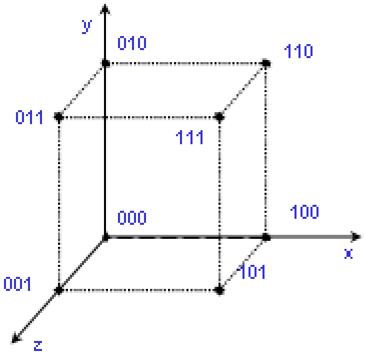
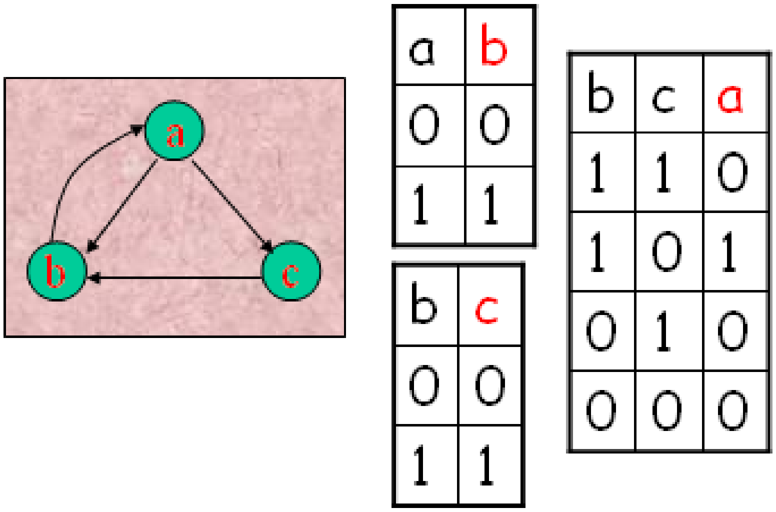


Figure 10.7: The state space of 3 states.

Examples

Figure 10.8 gives an example of a simple boolean network and associated truth tables. This example shows a network of three nodes - a, b and c. As one can see, the expression of c directly depends on the expression of b, which in turn directly depends on a. Note that b influences more than one node, a and c (**"pleiotropic regulation"**), and that a is influenced by more than one node (**"multigenic regulated"**).



a	b
0	0
1	1

b	c
0	0
1	1

b	c	a
1	1	0
1	0	1
0	1	0
0	0	0

Figure 10.8: source: [10].A sample boolean network. The functions on the node are described by the tables: the right column is the value of the regulatee at the next stage (depends on the values of the other columns).

The assignment of values to nodes fully describes the *state* of the model at any given time. The change of model state over time is fully defined by the functions in F. Initial assignment of values uniquely defines the model state at the next step and consequently, on all the future steps. Thus, the network evolution is represented by its *trajectory*.

Figure 10.9 shows two such trajectories for the sample network. Since the number of possible states is finite, all trajectories eventually end up in single *steady state*, or a cycle of steady states.

Definition An *attractor* of a trajectory is a single steady state, or a cycle at the end of the trajectory. *The basin of attraction* for a specific attractor is the set of all trajectories leading to it.

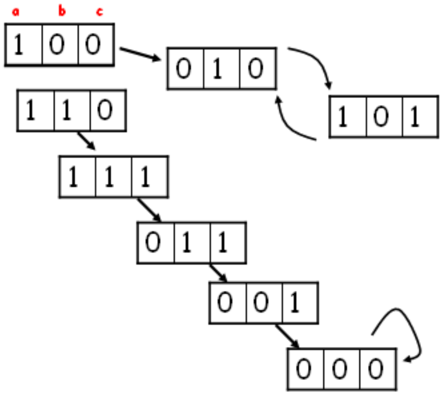


Figure 10.9: source: [10].States trajectories. The upper trajectory has a cycle of 2 steady states, while the lower trajectory ends in a single steady state. The basin of attraction of the upper trajectory is $[1,0,0],[0,1,0],[1,0,1]$, and the attractor of the upper trajectory is $[0,1,0],[1,0,1]$

One or more attractors are possible. The network in our example has two attractors - one is the steady state $(0,0,0)$, and the other is a cycle $(0,1,0) \leftrightarrow (1,0,1)$. The attractors are reached when $t \rightarrow \infty$. In a finite boolean network, one of the attractors is reached in a finite time.

States in genetic networks are often characterized by *stability* - ”slight” changes in value of a few nodes do not change the attractor. Biological systems are often *redundant* to ensure that the system stays stable and retains its function even in the presence of local anomalies. For example, there may be two proteins, or even two different networks with the same function, to backup each other.

10.1.9 Ensembles of Networks

We defined above what an *"ensemble of networks"* is, and which properties it possesses. However, each network has its own dynamics. The main features of the model, attractors and basins, are determined by the degree of connectivity in each network. A degree of connectivity k means that the in-degree of each node is exactly k.

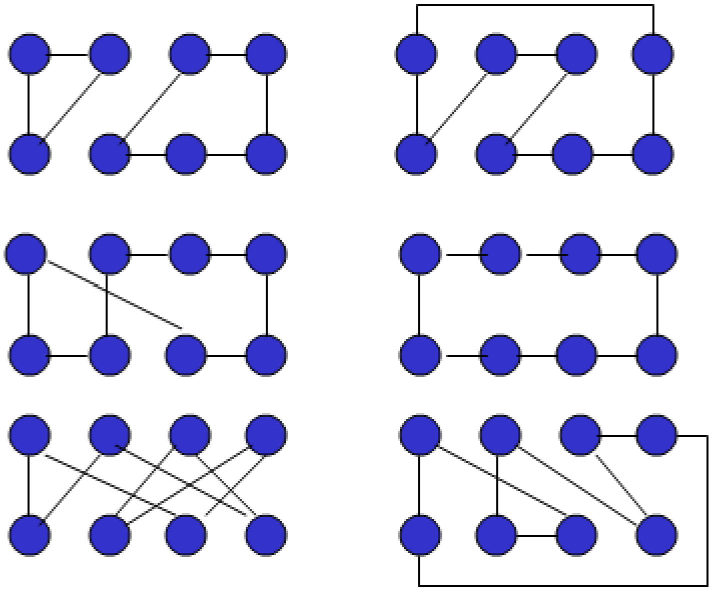


Figure 10.10: An ensemble of random networks with ($k = 2$). Note that every node in every network has degree 2.

High In-Degree

In the case that k is as high as $N - 1$:

- $X(t+1)$ is completely uncorrelated to $X(t)$, the output associated to each input set is random. There is no correlation between outputs corresponding to two inputs which differ even by a single bit. The system is chaotic and the homeostatic stability is very low, nearby initial states go to different attractors, and changing one input function completely destroys the basin structure.

- The number of attractors, about N/e, is very small compared to the 2^N possible states.

- The cycles are huge, period size is around $2^{0.5N}$.

For example, for $N = 100,000$ we get $10^{30,000}$ states, only 37,000 attractors and cycles are as long as $10^{15,000}$.

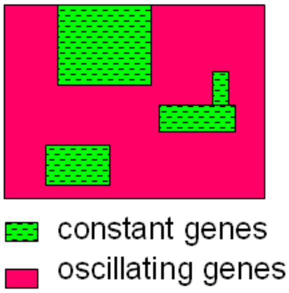


Figure 10.11: A 2-dimensional lattice view of a generic network, i.e., every cell in the lattice represents a gene. It can be seen, that when the in-degree is high, most of the genes are oscillating, that is, their state changes very frequently, and only few genes reach a constant state. Furthermore, the oscillating genes form a giant component, instead of being scattered all over the lattice.

Low In-Degree

In the case of $k = 2$:

- Basins are regular: nearby initial states usually reach the same attractor, high homeostatic stability, spontaneous order, even though inputs and functions are completely random.

- The number of attractors is relatively high - about $N^{1/2}$.

- Average cycle length is $N^{1/2}$.

Phase Transition

For a k-input boolean function, define $P = \max\{\text{no. 1-outputs, no. 0-outputs}\}/2^k$. It's obvious that $0.5 \leq P \leq 1$. For $P \approx 0.5$, the function is chaotic. For $P \approx 1$, the function is almost constant. In order to control the phase transition for different values of k by changing P, for example, by using canalizing functions, a boolean function where there is at least one value of one of the inputs that uniquely determines the output, irrespective of the others (eg. AND, OR).

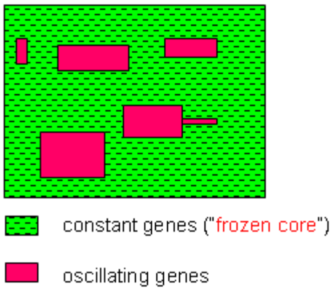


Figure 10.12: A 2-dimensional lattice view of a generic network with low in-degree. One can see that the effect is opposite to that observed in Figure 10.11 - most of the genes are constant, forming a giant component, while only few genes oscillate.

10.1.10 Concluding Remarks about Kauffman's Model

A possible explanation of the model

The model is consistent with experimental observations over many different phyla. A ratio that was observed is that the number of cell types is approximately the number of different cycles which is approximately the number of genes$^{0.5}$. A possible explanation is that a different cell start position will develop different types of cell. Another ratio that was observed is the length of cell life is approximately the length of the cycle in the graph.

Summery

Kauffman's model is a highly idealized representation of real genetic networks, due to the following reasons:

- The relation between genes are discrete (boolean) rather than continuous.
- The network status at time $t + 1$ depends only on its status at time t.
- Chemicals are not taken into account.
- Regulatory proteins are assumed to be synthesized very fast with respect to the regulation process itself.
- Synchronous activation may introduce "spurious cycles" in boolean dynamical systems.
- Fixed in-degree k is assumed for all genes.

However, Kauffman's model allows us to address issues which would otherwise be neglected, and to develop an appropriate language in which we can formulate key questions, such as:

- The importance of attractors in determining the properties of genetic networks.
- Robustness and basins of attraction.
- The importance of the average degree of connectivity.

Kauffman's model also allows us to examine in a new way the interplay between selection and self-organization. Moreover, it demonstrates the importance of studying ensembles of networks to gain insight about their generic properties.

10.2 Identification of Gene Regulatory Networks by Gene Disruptions and Overexpressions

10.2.1 Preface

This section is based on the article of Akutsu et al. [3]. Almost all proofs and all figures were take from this paper. In this section we show how to identify a gene regulatory network from data obtained by multiple gene perturbations (disruptions and overexpressions) taking into account the number of experiments and the complexity of experiments. An experiment consists of parallel gene perturbations and their total number is the complexity of an experiment.

10.2.2 Model Description and Definitions

We define the gene regulatory network as in the previous section. We further assume that it satisfies the following conditions:

1. When the boolean function f_v assigned to v has k inputs, k input lines (directed edges) come from k distinct nodes $u_1, ..., u_k$ other then v.

2. For each $i = 1, ..., k$ there exists an input $(a_1, .., a_k) \in \{0, 1\}^k$ with $f_v(a_1, ..., a_k) \neq f_v(a_1, .., \bar{a}_i, ..., a_k)$ where $\bar{a}_i$ is a complement bit of a_i.

3. A node v with no inputs has a constant value (0 or 1).

Definition The *state* of a gene v is active (inactive) if the value of v is 1 (0).

Definition The node v is called $AND(OR)$ node if the value of $f_v(a_1, ..., a_k)$ is determined by the formula $\ell(u_1) \wedge \ell(u_2) \wedge ... \wedge \ell(u_k)$ $(\ell(u_1) \vee \ell(u_2) \vee ... \vee \ell(u_k))$, where $\ell(u_i)$ is either u_i or $\neg u_i$.

Definition An edge (u, v_i) is called an *activation edge (inactivation edge)* if $\ell(u_i)$ is a positive literal (negative literal).

For a gene v, *a disruption* of v forces v to be inactive and *overexpression* of v forces v to be active. Let $x_1, ..., x_p, y_1, ..., y_q$ be mutually distinct genes of G. An *experiment* with gene overexpressions $x_1, ..., x_p$ and gene disruptions $y_1, ..., y_q$ is denoted by $e = \langle x_1, ..., x_p$, $\neg y_1, ..., \neg y_q \rangle$. The *cost* of e is defined as $p + q$. Three cases of gene expression conditions (normal, disruption of gene A, overexpression of gene B) are presented in figure 10.14.

Let us define the nodes with fixed values given *experiment* e:

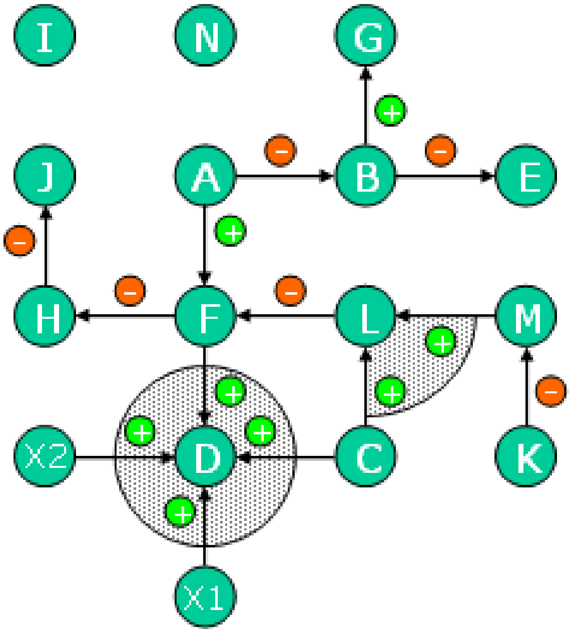
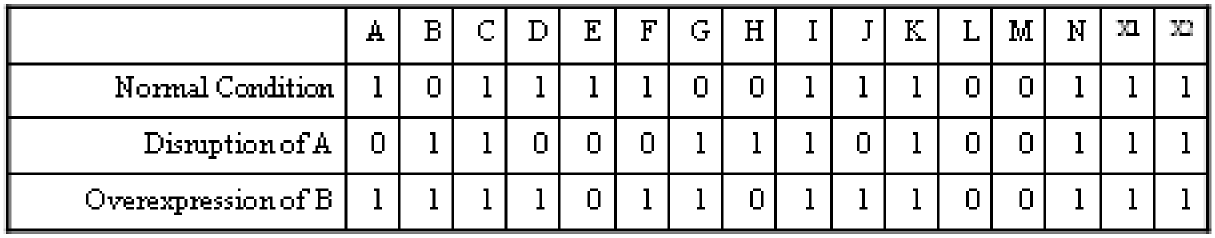


Figure 10.13: Source: [3]. Example of a gene regulatory network with 16 genes ($\oplus$ means ”activation” and $\ominus$ means ”deactivation” of the gene). Gene F is *activated* by gene A and is also *inactivated* by gene L ($f_F(A, L) = l(A) \wedge \neg l(L)$). Gene D is expressed if all its predecessors $C, F, X1, X2$ are expressed (AND - node).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	X1	X2
Normal Condition	1	0	1	1	1	1	0	0	1	1	1	0	0	1	1	1
Disruption of A	0	1	1	0	0	0	1	1	1	0	1	0	0	1	1	1
Overexpression of B	1	1	1	1	0	1	1	0	1	1	1	0	0	1	1	1

Figure 10.14: Source: [3]. Gene expressions by disruption and overexpression from the gene regulatory network of Figure 10.13 (0 - the gene is not expressed , 1 - the gene is expressed).

Definition The node v is said to be *invariant* if it satisfies one of the following conditions:

- v belongs to e, i.e., v is disrupted or overexpressed in e.
- v has in-degree 0.
- v depends only on invariant nodes.

We now define different types of states of gene regulatory network G:

1. A *global state* of G is a mapping $\psi : V \to \{0, 1\}$. The global states of the genes need not be consistent with the gene regulation rules.
2. The global state ψ of G is *stable* under experiment $e = \langle x_1, ..., x_p \ , \neg y_1, ..., \neg y_q\rangle$ if $\psi(x_i) = 1 \ (i = 1, ..., p)$, $\psi(y_j) = 0 \ (j = 1, ..., q)$ and it is consistent with all gene regulation rules, i.e., for each node v with inputs $u_1, ..., u_k$, $\psi(v) = f_v(\psi(u_1), ..., \psi(u_k))$. Otherwise, it is called *unstable*.
3. The global state ψ of G is an *observed global state* under experiment $e = \langle x_1, ..., x_p$, $\neg y_1, ..., \neg y_q\rangle$ if it satisfies all gene regulation rules for invariant nodes.
4. The observed global state ψ of G is a *native global state* when no perturbations are made ($e = \langle\rangle$).

We shall now prove upper and lower bounds for the number of experiments required for identifying a gene regulatory network with n genes, depending on the in-degree constraint and acyclicity. Table 10.1 summarizes the results. Computationally the running time of all algorithms when the in-degree is bounded is polynomial.

10.2.3 Upper and Lower Bounds on the Number of Experiments

We first show that an exponential number of experiments are required in the worst case.

Proposition 10.1 $\Theta(2^{n-1})$ *experiments must be performed in order to identify a gene regulatory network in the worst case.*

Proof: Consider a boolean function of $(n-1)$ variables $f(x_1, x_2, .., x_{n-1})$ which is assigned to the node x_n. There are $2^{2^{n-1}}$ possible boolean functions of $(n-1)$ variables. Hence we can identify this function by examining 2^{n-1} assignments and less examinations will not suffice (we get one output bit per experiment). ■

Proposition 10.2 $n2^{n-1}$ *experiments always suffice in order to identify a gene regulatory network.*

Constraints	**Lower bounds**	**Upper bounds**
None	$\Omega(2^{n-1})$	$O(2^{n-1})$
In-degree $\leq D$	$\Omega(n^D)$	$O(n^{2D})$
In-degree $\leq D$ All genes are AND-nodes (OR-nodes)	$\Omega(n^D)$	$O(n^{D+1})$
In-degree $\leq D$ Acyclic	$\Omega(n^D)$	$O(n^D)$
In-degree ≤ 2 All genes are AND-nodes (OR-nodes). No inactivation edges.	$\Omega(n^2)$	$O(n^2)$

Table 10.1: Source: [3]. Bounds on the number of experiments needed for reconstruction (n - number of genes, D - maximum in-degree). As seen from the table, forcing more constrains on the possible network topologies can improve experimental complexity significantly. The cases of acyclic topologies and restricted monotone logic (AND/OR gates only) are simpler mathematically but have no biological motivation.

Proof: For each node 2^{n-1} experiments are sufficient to identify its Boolean function by Proposition 10.1. Hence $n2^{n-1}$ experiments suffice in order to identify the whole network. ■

Theorem 10.3 *An exponential number of experiments are necessary and sufficient for the identification of a gene regulatory network.*

10.2.4 Bounded In-degree Case With Bounded Cost

Since an exponential lower bound was proved in the general case, we consider a special but practical case, in which the maximal in-degree is bounded by a constant D. First, we consider the case $D = 2$.

Proposition 10.4 $\Omega(n^2)$ *experiments are necessary for identification even if the maximum in-degree is 2 and all nodes are AND nodes, where we assume that the maximum cost is bounded by a fixed constant* C.

Proof: First, consider the case of $C = 2$. Assume that $\neg x \wedge \neg y \rightarrow z$ is assigned to z and all other nodes have in-degree 0. Among all experiments only $(\neg x, \neg y)$ can activate z. Therefore, we must test $\Omega(n^2)$ pairs of nodes in order to find (x, y).

Next, we consider the case of $C = 3$ with the same function $\neg x \wedge \neg y \rightarrow z$. If we disrupt or overexpress u, v, w such that $x \notin \{u, v, w\}$ or $y \notin \{u, v, w\}$, we can only learn that $(u, v), (u, w), (v, w)$ are different from (x, y). Since there are $\Theta(n^3)$ triplets and only $\Theta(n)$ triplets can include $\{x, y\}$, $\Theta(n^2)$ triplets must be examined in the worst case (each experiment removes at most a constant number of pairs out of the $\Theta(n^2)$ possible ones).

For cases of $C > 3$, similar arguments work: suppose $C = k > 3$, if we disrupt and/or everexpress $u_1, ..., u_k$ such that $x \notin \{u_1, ..., u_k\}$ or $y \notin \{u_1, ..., u_k\}$, we can only know that $\frac{k!}{2! \cdot (k-2)!}$ pairs are different from (x, y). Since there are $\Theta(n^k)$ k-mers and only $\Theta(n^{k-2})$ k-mers can include $\{x, y\}$, $\Theta(n^2)$ triplets must be examined in the worst case (each experiment removes at most a constant number of pairs out of the $\Theta(n^2)$ possible ones). ■

If C is not bounded, the above proposition does not hold. It is possible to identify the above pair (x, y) by $O(\log(n))$ experiments of maximum cost n, using a strategy based on binary search. Although this strategy might be generalized for other cases, we do not investigate it because experiments with high cost are not realistic. (The cells simply die if they are heavily mutated.)

Next, we consider the upper bound.

Proposition 10.5 $O(n^4)$ *experiments with maximum cost 4 are sufficient for identification if the maximum in-degree is 2.*

Proof: We assume (w.l.o.g.) that all nodes are of in-degree 2 since identification of nodes of in-degree 1 or 0 is easier. Let c be any node of V. We examine all assignments to all quadruplets $\{a, b, x, y\}$ with $c \notin \{a, b, x, y\}$. The boolean function $g(a, b)$ is assigned to c (i.e., $f_c \equiv g$) if and only if $c \equiv g(a, b)$ for any assignment to $\{a, b, x, y\}$, where $c \equiv g(a, b)$ means that the *state* of c equals to $g(a, b)$. The 'only if' part is trivial. We shall prove the 'if' part. Suppose that $g(a, b)$ is not assigned to c, i.e., $f_c = h(a, b)$ and $h(a, b) \neq g(a, b)$. Clearly, $c \equiv g(a, b)$ does not hold. Next, consider the case where $h(p, q)$ is assigned to c where h may be equal to g and $\{p, q\} \cap \{a, b\} = \emptyset$. In this case, c takes both 1 and 0 by changing assignments to $\{p, q\}$ even if the assignment to $\{a, b\}$ is fixed. Therefore, $c \equiv g(a, b)$ does not hold. In the case remaining $\{p, q\} \cap \{a, b\} \neq \emptyset$. Suppose $f_c \equiv h(p, b)$ and $a \neq p$. Then there is a value of b so that $h(0, b) \neq h(1, b)$, but then $f_c(a, b, p = 0, y) \neq f_c(a, b, p = 1, y)$ and $c \equiv g(a, b)$ does not hold again. Since all assignments to all quadruplets are examined, in total $0(n^4)$ experiments are sufficient. ■

The above property holds even for an *unstable* graph because c is consistent under any experiment on $\{a, b, x, y\}$ if $f_c \equiv g(a, b)$.

Theorem 10.6 *$O(n^{2D})$ experiments with maximal cost $2D$ are sufficient for the identification of a gene regulatory network of bounded in-degree D. On the other hand, $\Omega(n^D)$ experiments are necessarily in the worst case if the cost of each experiment is bounded by a constant.*

10.2.5 Efficient Strategies for Special Cases

In this section we consider the case where the network consists of AND and/or OR nodes. In this case we assume that any AND (resp. OR) node c is *inactive* (resp. *active*) if at least one literal appearing in the boolean function assigned to c is forced to 0 (resp. 1) by disruption or overexpression of the gene corresponding to the literal. The above assumption is biologically reasonable even when the network contains inconsistent nodes.

Theorem 10.7 *A gene regulatory network which consists of AND and/or OR nodes and has maximum in-degree D can be identified by $O(n^{D+1})$ experiments.*

Proof: Here we only show strategy for a network that consists of AND nodes of in-degree 2. It can be generalized though, to the other cases. We examine all assignments to all triplets $\{a, b, x\}$ with $c \notin \{a, b, x\}$. The function $g(a, b)$ is assigned to c (i.e., $f_c = g$) if and only if $c \equiv g(a, b)$ for any assignment to $\{a, b, x\}$. Following the proof in Proposition 10.5, we only have to consider the case that $h(p, q)$ is assigned to c and $\{p, q\} \cap \{a, b\} = \emptyset$. Consider an assignment to $\{a, b, p\}$ for which $g(a, b) = 1$. If c is not *active* we can conclude that $c \equiv g(a, b)$ does not hold. If c is *active*, we can inactivate c by changing the assignment to p since only one assignment to $\{p, q\}$ can activate c. Thus , $c \equiv g(a, b)$ does not hold. Therefore, the above property holds and $O(n^3)$ experiments are sufficient in total. ■

Next, we consider the acyclic case for which we obtain an optimal bound.

Definition A set of nodes $\{x_1, x_2, ..., x_k\}$ has *influence* on y if there exist two experiments e_1 and e_2 on $\{x_1, x_2, ..., x_k\}$ such that e_1 activates y and e_2 inactivates y.

Definition A set of nodes $\{x_1, x_2, ..., x_k\}$ has *influence* on $\{y_1, y_2, ..., y_p\}$ if $\{x_1, x_2, ..., x_k\}$ has influence on at least one of $\{y_1, y_2, ..., y_p\}$.

Definition A set of nodes $\{x_1, x_2, ..., x_k\}$ has strong *influence* on y if there exist two experiments e_1 and e_2 on $\{x_1, x_2, ..., x_k\}$ such that e_1 activates y and e_2 inactivates y, and e_1 differs from e_2 only on a single x_i.

The above definitions are invalid if the network is unstable (i.e., has an inconsistent node) or has multiple stable states. Henceforth , we assume that the network cannot have inconsistent nodes except ones that are disrupted or overexpressed. Moreover, for stable networks, we make a biologically reasonable assumption that a set of nodes $\{x_1, x_2, ..., x_k\}$ does not have influence on a node to which there is no direct path from any of $\{x_1, x_2, ..., x_k\}$.

Theorem 10.8 *An acyclic gene regulatory network of maximum in-degree D can be identified by $\Theta(n^D)$ experiments.*

Proof: The lower bound directly follows from Proposition 10.4 and Theorem 10.6. We prove the upper bound only for $D = 2$. Other cases can be proved in similar way. Moreover, we only show the strategy for a node with $a \wedge b \rightarrow c$ although it can be generalized to other types of nodes. We assume (w.l.o.g.) that all nodes are of in-degree 2 as in Proposition 10.5. Let P be a set of pairs (x, y) satisfying the following conditions: c is *active* under $\langle x, y \rangle$, and c is *inactive* under the other assignments to (x, y). Then $a \wedge b \rightarrow c$ if and only if $(a, b) \in P$ and (a, b) does not have influence on any other pair $(x, y) \in P$. If $a \wedge b \rightarrow c$, then $(a, b) \in P$ must hold. Moreover, (a, b) does not have influence on any other pair in P since the network is acyclic. Conversely, if $a \wedge b \rightarrow c$ does not hold, then $(a, b) \notin P$ or (a, b) has influence on at least one node x, such that there is an edge from x to c. Therefore, we can identify the network by $O(n^2)$ experiments with maximum cost 2. ■

For cyclic networks with in-degree, 2 experiments of cost 2 do not suffice. It is possible to identify such network in some cases in $O(n^D)$ experiments. The strategy is based on detection of strongly connected components.

10.2.6 Related Problems: Consistency and Stability of Networks

Along with the identification of the gene regulatory network, there exist several important problems. Here we observe two of them.

1. The consistency problem: given a network $G'(V', F')$, check whether or not this network coincides with an underlying gene regulatory network $G(V, F)$, that is not given explicitly.

   Theorem 10.9 *Exponential number of experiments are necessary and sufficient to check the consistency of a given gene regulatory network.*

2. The stability problem: given a network $G(V, F)$, check whether or not it is stable (in a native state), i.e., there is a global state consistent with all gene regulation rules.

   Theorem 10.10 *Testing the stability of a given gene regulatory network under an experiment is* NP-complete*.*

Bibliography

[1] http://www.expasy.ch/ch2d/.

[2] http://www.uib.no/aasland/two-hybrid.html.

[3] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano. Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 695–702, San Francisco, California, 25–27 January 1998.

[4] P. D'haeseleer and S. Fuhrman. Gene network inference using a linear, additive regulation model. *Bioinformatics*, 2000.

[5] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mRNA expression levels during CNS development and injury. In *Pacific Symposium on Biocomputing*, pages 41–52, Hawaii, Hawaii, 1999. World Scientific Publishing Co.

[6] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J. Computational Biology*, 7(3):601–620, Nov 1998.

[7] S.A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, 1993.

[8] S.A. Kauffman. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity.* Oxford University Press, 1995.

[9] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing*, pages 18–29, Maui, Hawaii, 1998. World Scientific Publishing Co.

[10] R. Somogyi and C. Sniegoski. *Complexity*, 1:45–63, 1996.

[11] H. Zhu, J.F. Klemic, S. Chang, P. Bertone, A. Casamayor, K.G. Klemic, D. Smith, M. Gerstein, M.A. Reed, and M. Snyder. Analysis of yeast protein kinases using protein chips. *Nature Genetics*, 26:283–289, 2000.