Session on Gene Expression and Genetic Networks

Pacific Symposium on Biocomputing, 1999

Hawaii , January 4-9, 1999

Tutorial:

Gene Expression Data Analysis and Modeling

Patrik D'haeseleer, Shoudan Liang and Roland Somogyi

Patrik D'haeseleer
University of New Mexico, Dept. Computer Science, Albuquerque, NM 87131
(patrik@cs.unm.edu)

Shoudan Liang
SETI Institute, NASA Ames Research Center, Moffett Field, CA 94035
(sliang@mail.arc.nasa.gov)
Also consulting for Incyte Pharmaceuticals Inc., 3174 Porter Dr., Palo Alto, CA 94304

Roland Somogyi
Incyte Pharmaceuticals Inc., 3174 Porter Dr., Palo Alto, CA 94304
(rsomogyi@incyte.com)

Introduction

The traditional approach to research in Molecular Biology has been an inherently local one, examining and collecting data on a single gene, a single protein or a single reaction at a time. This is, of course, the classical reductionist stance: to understand the whole, one must first understand the parts. Over the years, this approach has led to remarkable achievements, allowing us to make highly accurate biochemical models of such favorites as bacteriophage Lambda.

However, with the advent of the "Age of Genomics" an entirely new class of data is emerging. To date, analysis of this large scale data has consisted of little more than descriptions of how many genes were previously unknown, which genes are over- or underexpressed under certain circumstances, etc. Of course, such data is a valuable resource for researchers who are focusing on individual genes. But can we really expect to construct a detailed biochemical model of, say, an entire yeast cell with some 6000 genes (only about 1000 of which were defined before sequencing started, and about 50% of which are clearly related to other known genes), by analyzing each gene and determining all the binding and reaction constants one by one? Likewise, from the perspective of drug target identification for human disease, we cannot realistically hope to characterize all the relevant molecular interactions one-by-one as a requirement for building a predictive disease model.

There is a need for methods that can handle this data in a global fashion, and that can analyze such large systems at some intermediate level, without going all the way down to the exact biochemical reactions. At the very least, such an analysis could help guide the traditional pharmacological and biochemical approaches towards those genes most worthy of attention among the thousands of newly discovered genes. Ideally, a sufficiently predictive and explanatory model at an intermediate level could obviate the need for an exact understanding of the system at the biochemical level.

Large scale gene expression mapping is motivated by the premise that the information on the functional state of an organism is largely determined by the information on gene expression (based on the central dogma). This process may be conceptualized as a genetic feedback network, in which information flows from gene activity patterns through a cascade of inter- and intracellular signaling functions back to the regulation of gene expression. Gene sequence information in cis regions (regulatory inputs) and protein coding regions (regulatory outputs; determines biomolecular dynamics) is expanded into spatio-temporal structures defining the organism. Some principles of this behavior may be captured by computational models, such as Boolean networks.

In order to draw meaningful inferences from gene expression data, it is important that each gene is surveyed under several different conditions, preferably in the form of expression time series. Such data sets may be analyzed using a range of methods with increasing depth of inference, such as cluster analysis, correlation analysis, and determination of mutual information content. Abstract computational models may serve as a test bed for the development of these inference techniques. Only in such models can the dynamic behavior of many elements (trajectories, attractors) be unequivocally linked to a selected network architecture (wiring and rules). Beyond cluster analysis lies the more ambitious realm of genetic network inference: complete reverse engineering of the underlying regulatory interactions from the expression data, either using idealized models such as Boolean networks, or using more realistic continuous models.

1. Introduction to principles of network behavior in the Boolean network model

1.1. Multigenic & pleiotropic regulation: the basis of genetic networks

From a strictly reductionist viewpoint, we may begin by asking "which gene underlies this disease?" or "with which molecule does this protein interact?". The resulting investigations have shown us that more often than not, several genes contribute to a disease, and that molecules interact with more than one partner (Fig.1). The challenge now lies in discovering what the significant connections are in these regulatory networks, and which abstract principles underlie the network architecture and dynamics allowing it to function in reliable way.

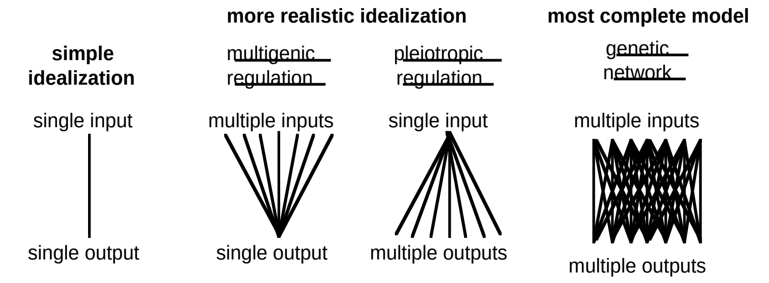
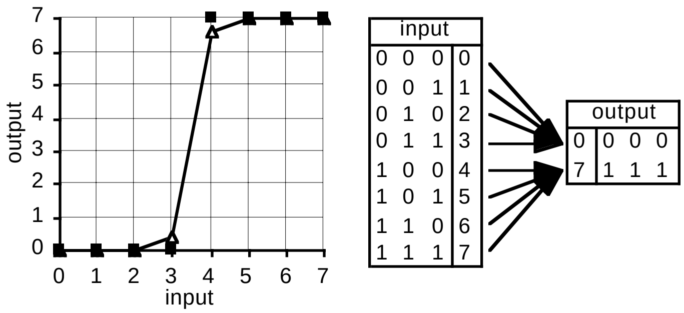


Figure1. From pair-wise interactions to networks. Somogyi R, Fuhrman S, Askenazi M, Wuensche A (1997) The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. Nonlinear Analysis, Proc. of Second World Cong. of Nonlinear Analysts (WCNA96), 30(3):1815-1824.

1.2. A simple, binary conceptualization of a biomolecular network

Perhaps a model based on idealized, elementary mechanisms can illustrate the nature of complex behavior. Boolean networks constitute such a model:

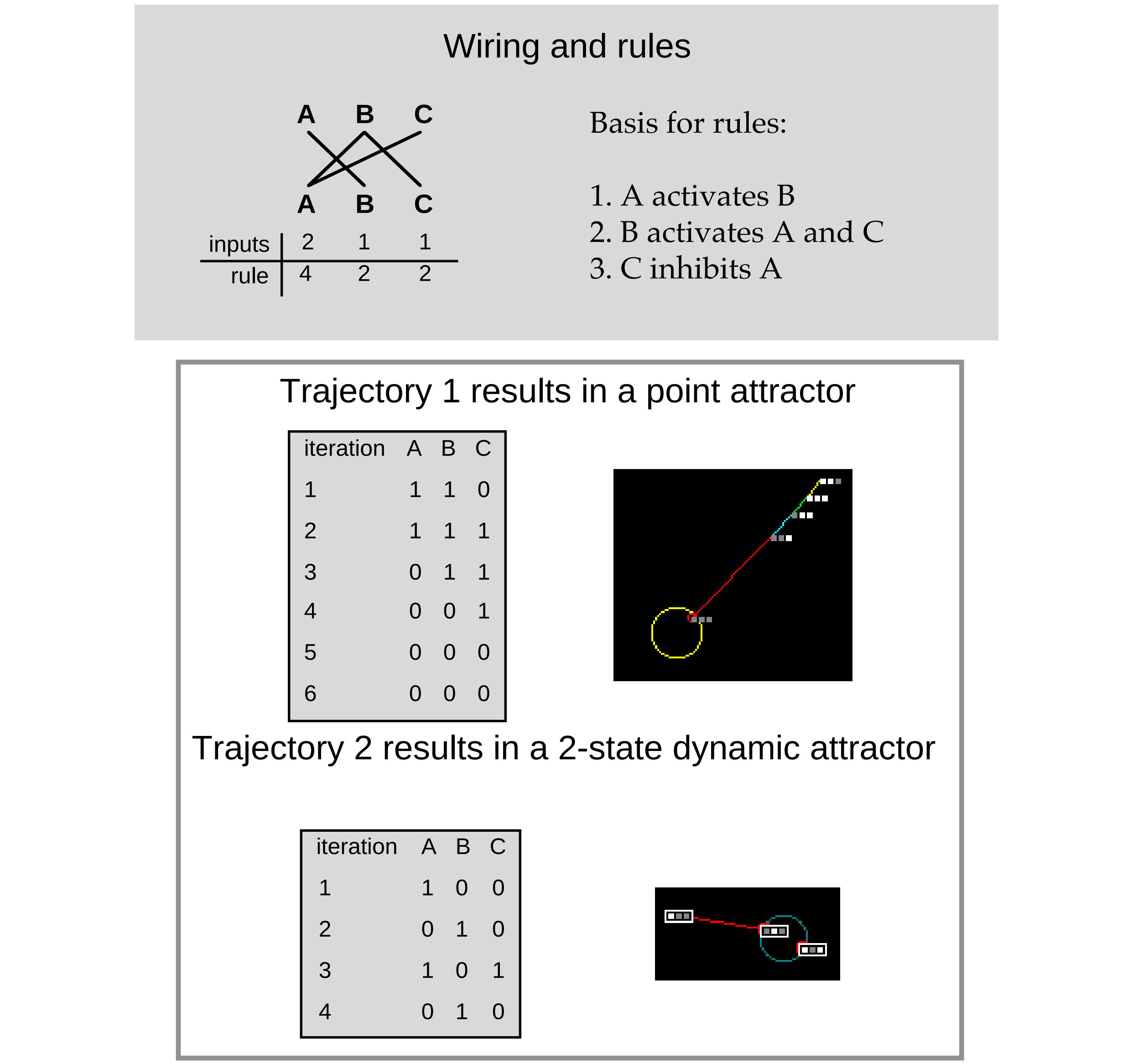
- Each gene may receive one or several inputs from other genes or itself (Fig. 1).
- Assuming a highly cooperative, sigmoid input-output relationship, a gene can be modeled as a binary element (Fig. 2).
- The output (time=t+1) is computed from the input (time=t) according to logical or Boolean rules. Time is discrete, and all genes are updated simultaneously.



input			
0 0 0	0
0 0 1	1
0 1 0	2
0 1 1	3
1 0 0	4
1 0 1	5
1 1 0	6
1 1 1	7

output	
0	0 0 0
7	1 1 1

Figure 2. From continuous to discrete kinetics. Sigmoid interactions represent a form of data reduction, a "many to one mapping", which has profound implications on systems stability. Somogyi, R (1998) Many to One Mappings as a Basis for Life.Interjournal (http://rsb.info.nih.gov/mol-physiol/ICCS/ms/mappings.html)



Wiring and rules

inputs	2	1	1
rule	4	2	2

Basis for rules:

1. A activates B
2. B activates A and C
3. C inhibits A

Trajectory 1 results in a point attractor

iteration	A	B	C
1	1	1	0
2	1	1	1
3	0	1	1
4	0	0	1
5	0	0	0
6	0	0	0

Trajectory 2 results in a 2-state dynamic attractor

iteration	A	B	C
1	1	0	0
2	0	1	0
3	1	0	1
4	0	1	0

Figure 3. Network wiring diagram & rules. Somogyi & Sniegoski,1996; Complexity 1(6):45-63

1.3. Wiring and rules determine network dynamics

The dynamics of Boolean networks of any complexity are determined by the wiring and rules, or state-transition tables. The wiring diagram is shown in the top panel of Fig. 3. The lines connect the upper row of output elements (time=t) to the lower row of input elements (time=t+1). The no. of inputs and the pertaining decimal rule are shown underneath each wiring diagram. Time space patterns or trajectories (lower panels) can be directly calculated from the wiring and rules. The middle panel shows a point attractor, its basin of attraction including 5 states. The lower panel illustrates a 2-state dynamic attractor (repeating pattern), with a basin of 3 states. The basins of attraction cover all 8 possible states of the system.

1.4. Many states converge on one attractor

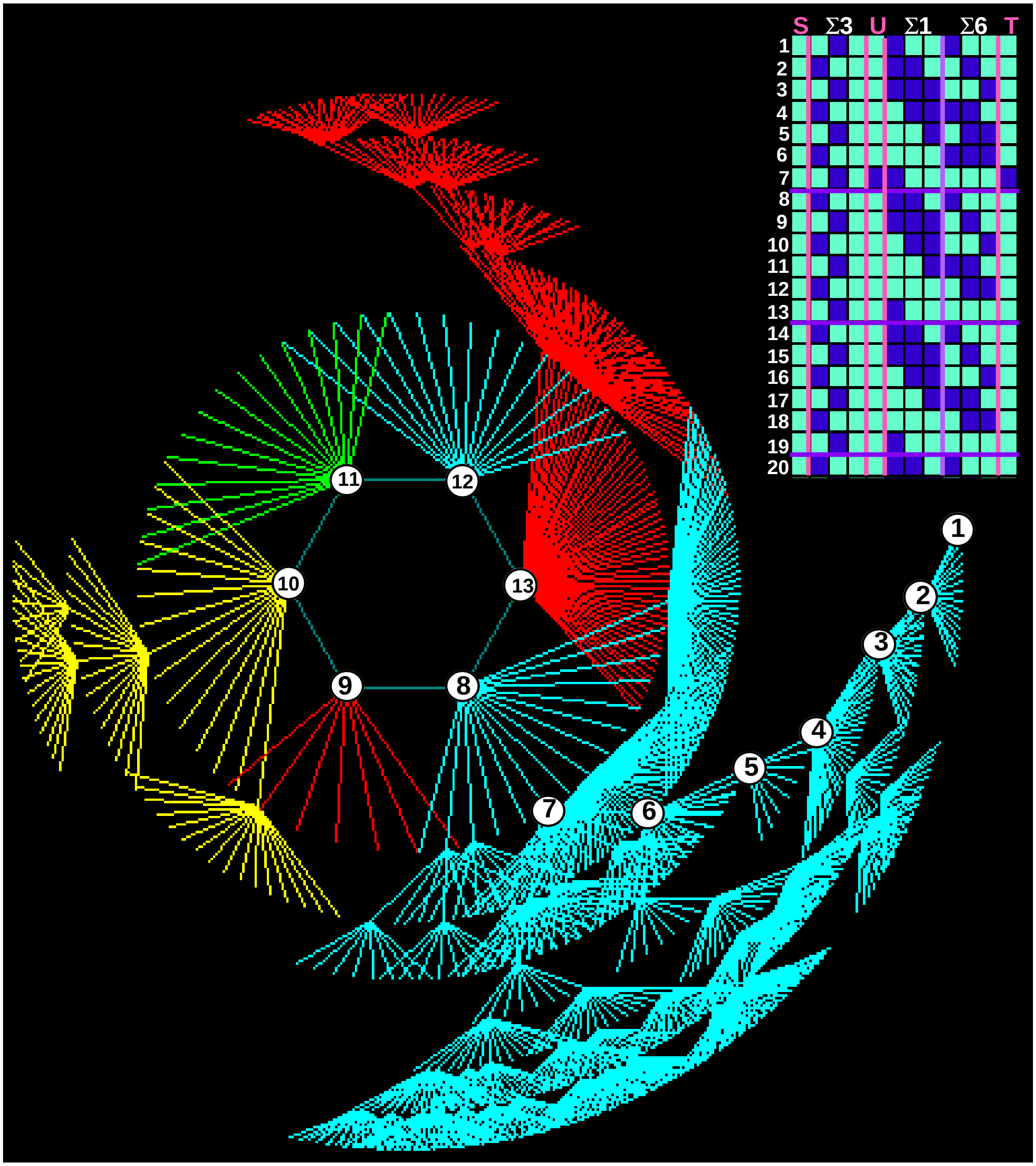


Figure 4. Basin of attraction of 12-gene Boolean genetic network model. Somogyi & Sniegoski,1996; Complexity 1(6):45-63. The model network trajectories and basins of attraction shown here were generated using the DDLAB software by Andy Wuensche, Ph.D. Andy Wuensche's web site: http://www.santafe.edu/~wuensch/

All Boolean network time series terminate in specific, repeating attractor patterns. These can be visualized as basin of attraction graphs (Fig. 4). The network trajectory (upper right panel) inexorably leads to a final state or state cycle - an attractor (center). Each state of the trajectory is shown as a point (labeled by its time step number). The labeled series of state transitions is one of many trajectories converging on the repeating, six-state attractor pattern. All of the centripetal trajectories leading to the attractor form the basin of attraction.

In summary, our perusal of Boolean genetic network models has given us some insight into network principles. Even given complex architecture of wiring and rules, these systems produce predictable and stable behavior: all trajectories are strictly determined, and many states converge on one attractor. This suggests for living systems that small perturbations altering a particular state of the system may have no effect on general system outcome. While distributed signaling systems may appear unwieldy from a superficial standpoint, they confer the stability and robustness that characterize evolved living systems.

1.5. Network terminology

	Architecture	
wiring	<->	biomolecular connections
rules (functions, codes)	<->	biomolecular interactions
	Dynamics	
state	<->	set of molecular activity values; e.g. gene expression, signaling molecules
state transition	<->	response to previous state
trajectory	<->	series of state transitions; e.g. differentiation, perturbation response
attractor	<->	final outcome; e.g. phenotype, cell type, chronic illness

2. Gene expression data

2.1. What's available?

2.1.1. mRNA levels

2.1.1.1. cDNA microarrays

Developed at Stanford University, the microarrays are glass slides on which cDNA has been deposited by high-speed robotic printing. They are ideally suited for expression analysis of up to 10,000 cDNA clones per array from EST (expressed sequence tag) sequencing projects (such as the private effort at Incyte Pharmaceuticals and the public Washington University project).

Microarrays measurements are carried out as differential hybridizations to minimize errors originating from cDNA spotting variability: mRNA from two different sources (e.g control and drug-treated), labeled with two different fluorescent dyes, is passed over the array at the same time. The fluorescence signal from each mRNA population is evaluated inependently, and then used to calculate the treated/control expression ratio.

Patrick Brown's lab at Stanford has used microarrays to measure gene expression levels for the entire yeast genome (approximately 6400 distinct cDNA sequences) during the diauxic shift (transition from sugar metabolism to ethanol metabolism), sporulation and the entire cell cycle. These data sets are publicly available. The Brown Lab also has an online guide to build your own arrayer and scanner. These microarrays have been commercialized by Incyte Pharmaceutical's Microarray Division (formerly Synteni). Incyte Gene Expression Microarrays (GEMs) are available with templates from human, rat, mouse, plant and microbial genomes.

2.1.1.2. *Oligonucleotide chips*

These chips, produced by Affymetrix, consist of small glass plates with thousands of short 20-mer oligonucleotide probes attached to their surface, The oligonucleotides are synthesized directly onto the surface using a combination of semiconductor-based photolithography and light-directed chemical synthesis. Due to the combinatorial nature of the process, very large numbers of mRNAs can be probed at the same time. However, manufacturing and reading of the chips requires expensive equipment. Current chips have over 65,000 different probes, with typically several probes for each mRNA.

Affymetrix currently manufactures GeneChips for 42,000 human genes and ESTs, 30,000 murine genes and ESTs, and 6,100 yeast ORFs (whole genome). Little data is publicly available, with the exception of a S. cerevisiae expression database generated in collaboration with Ron Davis' lab.

2.1.1.3. *RT-PCR*

To measure gene expression using RT-PCR (Reverse Transcriptase Polymerase Chain Reaction), the mRNA is first reverse-transcribed into cDNA, and the cDNA is then amplified to measurable levels using PCR. Using built-in calibration techniques, RT-PCR can achieve high accuracy coupled with an exceptional sensitivity of 10molecules/10µl assay volume and a dynamic range covering 6-8 orders of magnitude. The method does require PCR primers for all the genes of interest, and is not inherently parallel like the previous three, so automation is crucial to scale up.

Roland Somogyi has used this method to measure the expression levels of 112 genes at nine different time points during the development of rat cervical spinal cord, and 70 genes during development and following injury of the hippocampus. The former data set is publicly available, the second should be available soon.

2.1.1.4. *Serial Analysis of Gene Expression*

SAGE uses a very different technique for measuring mRNA levels. First, double stranded cDNA is created from the mRNA. A single 10 base pair (long enough to uniquely identify each gene) "sequence tag" is cut from a specific location in each cDNA. Then the sequence tags are concatenated into a long double stranded DNA which can then be amplified and sequenced. This method has two advantages: the mRNA sequence does not need to be known a priori—so it will also detect previously unknown genes—and it uses sequencing technology that many labs already have. The method is rather complex though, and requires a large amount of sequencing.

SAGE has been used to analyze the set of genes expressed during three different phases of the yeast cell cycle. SAGE has also been used to monitor the expression of at least 45,000 human genes in normal colon cells, colon tumors, colon cell lines, pancreatic tumors and pancreatic cell lines. Some of this data is available upon request.

2.1.2. Protein levels

Protein levels are much harder to quantify than mRNA levels. 2D-PAGE separates proteins on a two-dimensional sheet of gel, first in one direction based on their isoelectric point, and then in the other direction based on their molecular weight. The result is a two-dimensional image with a large number of protein "spots". The intensity of each spot is proportional to the amount of the specific protein present.

It is not a priori known which protein each spot represents, although the position of known proteins can be estimated. Also, new microsequencing and mass spectrometry techniques allow spots to be identified with proteins of which the sequence is known. The resolution of the spots may not be high enough to separate all proteins, and 2D gel results have been hard to reproduce, because of sensitivity to operating parameters and a host of possible artifacts. These problems have been somewhat alleviated lately by the use of highly standardized protocols and higher accuracy techniques.

There are several 2D gel databases for E. coli, yeast, Drosophila, rat, mouse, human, etc. One of the most important ones is the SWISS-2DPAGE database, containing a total of 518 entries from human, yeast, E. coli and Dictyostelium. . 2D-PAGE proteomics is currently being commercialized in a partnership between Incyte Pharmaceuticals and Oxford Glycosciences.

2.2. Data requirements for gene network inference

The purpose of this section is twofold: (1) to examine some of the difficulties and pitfalls associated with inferring gene networks from large-scale data; and (2) to provide some guidelines for experimentalists who are collecting such data with the intent of using it for genetic network inference.

2.2.1. The Curse of Dimensionality

Measuring more variables allows for a more exact model, but makes the correct model exponentially harder to find.

Our human intuition when faced with the task of modeling an unknown process is to observe as many parameters of the system as possible. This is clearly reflected in the current effort to measure the expression levels of more and more genes simultaneously, rather than to measure these expression levels as often as possible (which would require reusable or continuous measurement techniques).

However, in Machine Learning it is well known that the more variables one needs to model, the harder the modeling task becomes, because the size of the search space increases exponentially with the number of parameters of the model. This is often referred to as the Curse of Dimensionality.

Does this mean that our human intuition about modeling is wrong? Not necessarily. Although we humans do want to be able to look at as many variables of the problem as possible, we rather quickly select those we think are really important to the system, and simply ignore the others. Our reason for wanting to know all the variables is so we wouldn't miss any of the important ones, not so we could include all the non-important ones in our model. Similarly, in Machine Learning, careful selection of the input variables is crucial to get around the Curse of Dimensionality. Use of a priori information can also help narrow down the range of plausible models.

2.2.2. What are the important variables?

The state of a cell consists of all those parameters--both internally and externally--which determine its behavior. Following the Central Dogma of molecular biology, the activity of a cell is determined by which of its genes are being expressed or not. If a particular gene is being expressed, its DNA is transcribed into complementary messenger RNA (mRNA), which is then translated into the specific protein the gene codes for. We can measure the level of expression of each gene by measuring how many mRNA copies are present in the cell.

> *"The mRNA levels sensitively reflect the state of the cell, perhaps uniquely defining cell types, stages, and responses. To decipher the logic of gene regulation, we should aim to be able to monitor the expression level of all genes simultaneously ... " [Lander]*

This cartoon picture of the Central Dogma is of course highly incomplete. Apart from the classical DNA → mRNA → protein pathway, the genes in the DNA are themselves regulated by the presence or absence of certain proteins. Furthermore, many of the interactions going on in the cell occur entirely at the protein level, which can cause significant discrepancies between protein and mRNA levels. In a recent comparison of selected mRNA and protein abundances in human liver, a correlation of only 0.48 was observed between the two. Clearly, protein levels form an important part of the internal state of a cell.

In addition to mRNA and protein levels, one could imagine measuring a number of other parameters, including cell volume, growth rate, methylation states of DNA, phosphorylation state of proteins, localization of proteins and mRNA within the cell, ion levels, etc. One class of data which could be very useful is metabolite and nutrient levels.

For example, during the diauxic shift in yeast (transition from glucose metabolism to ethanol metabolism), one would of course need to measure glucose levels, but preferably also a number of other metabolites involved such as acetate, pyruvate, glycogen, trehalose, etc. Arkin et al. uses capillary zone electrophoresis to simultaneously measure eight of the small molecular species in an in vitro glycolysis reaction.

Currently, most studies trying to infer expression mechanisms from cell state data use mRNA levels, because they are the easiest to measure (especially with the new large-scale gene expression technologies). Large-scale protein measurements tends to be very incomplete (typically only measuring the highest abundancy proteins), but can be supplemented with more exact measurements of individual proteins which are known to play an important role. If most protein levels turn out to be exactly correlated with the corresponding mRNA levels, they can always be left out of the model. Similarly, when measuring gene expression data on a process involving metabolism (and which cellular process doesn't?), an effort should be made to quantitate the most important metabolite and nutrient levels.

2.2.3. Constraining the model

The space of models to be searched increases exponentially with the number of parameters of the model, and therefore with the number of variables. Narrowing down the range of plausible models by putting on extra constraints can simplify the search for the best model considerably. For example, constraining the genes to be regulated by no more than 7 other genes will drastically simplify the number of regulatory interactions we need to consider. Similarly, for Boolean networks, constraining the rules for each gene to be biologically plausible can significantly reduce the number of Boolean rules that match the data we have on the regulation of each gene.

Constraining the model by using a priori information about what is biologically known or plausible is probably the most important weapon we have to fight the Curse of Dimensionality! How precisely to include this information into the inference process is the true art of modeling.

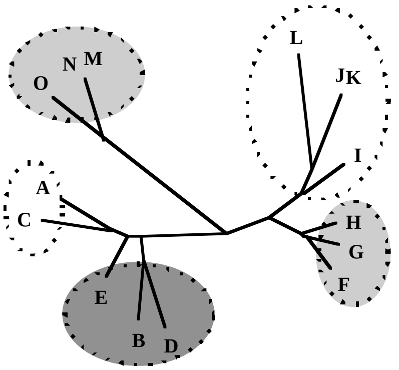
2.2.4. Number and variety of data points needed

The gene network inference techniques we will cover have one thing in common: they tend to be data-hungry. Measuring gene expression time series has the nice feature of yielding lots of data. However, all the data points in a single time series tend to be about a single dynamical process in the cell, and will be related to the surrounding time points. A data set of ten expression measurements under different environmental conditions, or with different mutations, will actually contain more information than a time series of ten data points on a single phenomenon. The advantage of the time series is that it can provide crucial insights in the dynamics of the process.

Both types of data, and multiple data sets of each, will likely be needed to unravel the regulatory interactions of the genes. Indeed, to correctly infer the regulation of a single gene, we need to observe the expression of that gene under many different combinations of expression levels of its regulatory inputs. This implies a wide variety of different environmental conditions and perturbations.

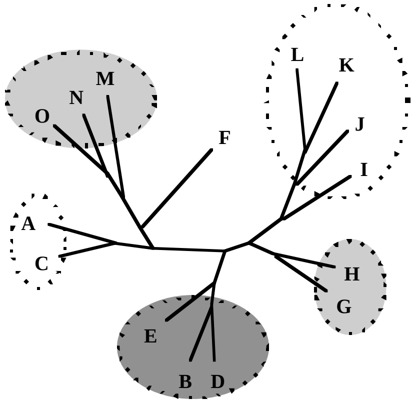
How much data points do we really need to infer a gene network on N genes? For a completely unconstrained, potentially fully connected Boolean network model, we would need to measure al possible 2^N input-output pairs. This is clearly inconceivable for realistic numbers of genes (30 genes would imply more than a billion data points needed). If we constrain the genes to have no more than K inputs from other genes, the number of (independent!) data points needed becomes proportional to log(N). John Hertz estimated K log(N/K) at PSB '98, but preliminary experimental results from Liang et al at PSB '98 and Akutsu et al at this Symposium, as well as calculations based on the probability that all the entries in the rule tables are uniquely specified after n independent input-output pairs, suggest the number of data points needed scales as 2^K log(N).

Wiring (Molecular Interaction) Clusters



gene	Boolean rule
A	F and H and J
B	G and H and J
C	F and H and I
D	G and H and I
E	H and I and J
F	I and J and K and L and (not G)
G	I and J and K and L and (not O)
H	I and J and K and L
I	J and K and L
J	K and L
K	K or L
L	L or M
M	N or O
N	N and O
O	N and O and (not E)

Trajectory (Gene Expression) Clusters



trajectory	I											II				III				IV		
time	1	2	3	4	5	6	7	8	9	10	1	2	3	4	1	2	3	4	1	2	3	4
A	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0
C	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0	1	0	0	0
D	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	0	0	1	1	0	0
E	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
G	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0
H	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0
I	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0
J	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5. Inference of shared control through cluster analysis. Somogyi R, Fuhrman S, Askenazi M, Wuensche A (1997) The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. Nonlinear Analysis, Proc. of Second World Cong. of Nonlinear Analysts (WCNA96), 30(3):1815-1824.

Similarly, for a fully connected linear or quasi-linear continuous model, we would need at the very least as many data points as genes. For models with restricted connectivity, we expect a similar improvement as for the Boolean case. From an information theory viewpoint, we retain more information by not quantizing the expression levels. Assuming a 15-20% quantitation error for RT-PCR, each measurement can give us up to 2-3 bits of information, whereas quantizing into Boolean values would give us only one bit. From cDNA microarrays and oligonucleotide chips, we can get 1-2 bits per measurement (assuming a 30-50% quantitation error).

3. Essential network analysis

3.1. Inference of shared control processes

In addition to studying models for the purpose of identifying organizational and dynamic principles, they provide an exploratory framework for the development of analytical tools. One of the major challenges in molecular signaling biology today lies in extracting functional relationships from e.g. gene expression time series. We will show how this can be accomplished using model networks.

3.1.1. Euclidean cluster analysis of a model network

As a first step toward, we need to address the classification of gene activity patterns:

- Similarities in gene expression patterns suggest shared control.
- Clustering gene expression patterns according to a heuristic distance measure is the first step toward constructing a wiring diagram.
- Euclidean distance as a measure for the difference between gene expression patterns: A gene expression pattern over n time points is a point in n-dimensional parameter space, therefore distance = $\sqrt{}\Sigma(a_i-b_i)^2$

In the example above (Fig. 5), we used Euclidean clustering to generate dendrograms of genes grouped according to shared inputs and shared dynamics . Note that the clustering pattern in the functional time series (lower panel) closely resembles the gene groupings according to wiring (upper panel). This analysis suggests that such clustering may be applied to biological activity data for the inference of shared control processes.

3.2. Principles of reverse engineering using Boolean networks

3.2.1. Biological information flow

As our experimental technologies become more sophisticated in sensitivity and throughput, we are generating vast amounts of information at all levels of biological organization. The challenge lies in inferring important functional relationships from these data. This problem is becoming acute as we are preparing to generate molecular activity data (e.g. mRNA and protein expression) for organisms in health and disease . This new perspective suggests to us to treat the organism as an information processing system. But how can we conceptualize information flow in living systems? There are several issues we need to address before applying the information concept in a non-trivial way.

- What is information?
- Can information be quantified?
- Can information measures be used in network analysis?

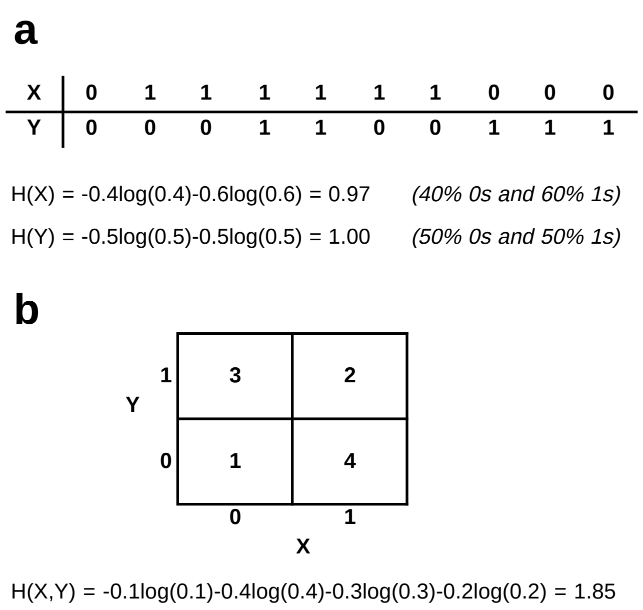
3.2.2. Information can be quantified: Shannon entropy (H)

From a mathematical-statistical standpoint, information can be quantified as the Shannon entropy (after Claude Shannon, the founder of information theory). The Shannon entropy (H) can be calculated from the probabilities of occurrences of individual or combined events as shown below:

$H(X)= - \Sigma\ p_x \log p_x$

$H(Y)= - \Sigma\ p_y \log p_y$

$H(X,Y) = - \Sigma\ p_{x,y} \log p_{x,y}$



X	0	1	1	1	1	1	1	0	0	0
Y	0	0	0	1	1	0	0	1	1	1

H(X) = -0.4log(0.4)-0.6log(0.6) = 0.97 *(40% 0s and 60% 1s)*

H(Y) = -0.5log(0.5)-0.5log(0.5) = 1.00 *(50% 0s and 50% 1s)*

Y \ X	0	1
1	3	2
0	1	4

H(X,Y) = -0.1log(0.1)-0.4log(0.4)-0.3log(0.3)-0.2log(0.2) = 1.85

Figure 6. Calculation of Shannon entropy from a series of observations. Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

The determination of H is illustrated in a numerical example (Fig. 6) . A single element is examined in a). Probabilities (p) are calculated from frequency of on/off values of X and Y. The distribution of value pairs is shown in b). H is calculated from the probability of co-occurrence of x, y values over all measurements.

3.2.3. The Shannon entropy is maximal if all states are equiprobable

The Shannon entropies for a 2-state information source (0 or 1) are graphed in Fig. 7. Since the sum of the state probabilities must be unity, p(1)=1-p(0) for 2 states.

3.2.4. Mutual information (M): the information (Shannon entropy) shared by non-independent elements

Figure 8 illustrates the relationships between the information content of individual and combined, non-independent information sources using Venn diagrams. Mutual information, M, is defined as the

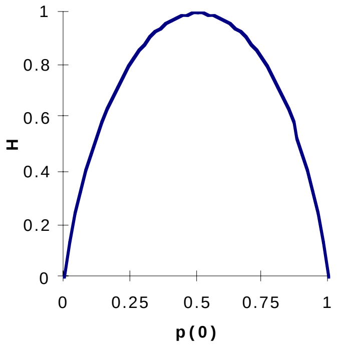


Figure 7. Shannon entropy vs. probability of event. Liang S, Fuhrman S, Somogyi R (1998) Pacific Symposium on Biocomputing 3:18-29.

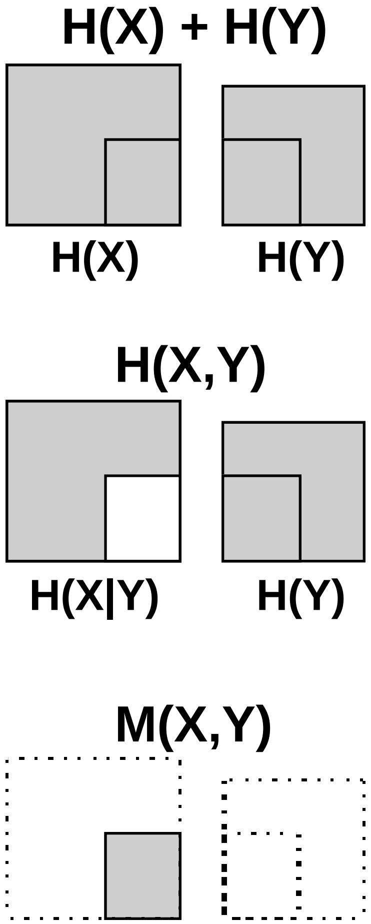
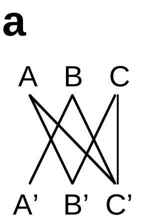


Figure 8 Venn diagrams of information relationships. Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

sum of the individual entropies minus the entropy of the co-occurrences:

M(X,Y) = H(X)+H(Y) - H(X,Y) .

In each case, add the shaded portions of both squares to determine one of the following: [H(X)+H(Y)], H(X,Y), and M(X,Y). The small corner rectangles represent information that X and Y have in common. H(Y) is shown smaller than H(X) and with the corner rectangle on the left instead of the right to indicate that X and Y are different, although they have some mutual information.



b

A' = B
B' = A or C
C' = (A and B) or (B and C) or (A and C)

c

input			output		
A	B	C	A'	B'	C'
0	0	0	0	0	0
0	0	1	0	1	0
0	1	0	1	0	0
0	1	1	1	1	1
1	0	0	0	1	0
1	0	1	0	1	1
1	1	0	1	1	1
1	1	1	1	1	1

Figure 9. Target Boolean network for reverse engineering. Liang S, Fuhrman S, Somogyi R (1998) Pacific Symposium on Biocomputing 3:18-29.

3.2.5. A candidate Boolean network for reverse engineering

Can measures of information be used to quantify causal relationships between elements fluctuating in a dynamic network? Using the example of Fig. 9, we will attempt to reconstruct the wiring diagram, a), and the Boolean rules, b), from the state transition table, c) (input column shows all states at time=t, outputs (prime) correspond to the matching states at time=t+1).

3.2.6. The principle behind REVEAL (REVerse Engineering ALgorithm)

The illustration in Fig. 10 (see next page) details the steps taken in REVEAL for the inference of functional connections and rules from the dynamics (i.e. state transition tables) of the Boolean network shown in Fig. 9. Hs and Ms are calculated from the time series or look-up tables according to the definitions (shaded). The wiring of the example Boolean network can be inferred from the state transition table using progressive M-analysis (left, odd steps). Once the inputs (wiring) to a gene are know, one can construct the rule table by matching the states of the inputs to those of the output from the state transition table (right, even steps).

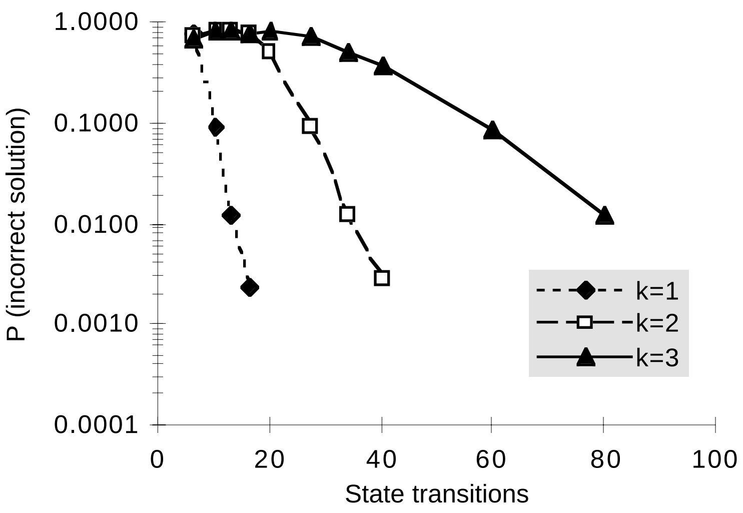


Figure 11 Calculation of Shannon entropy from a series of observations. Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

3.2.7. Inference from incomplete time series or state transition tables

REVEAL will quickly find a minimal solution for a Boolean network given any set of time series (Fig. 11). For n=50 (genes) and k=3 or less (number of inputs per gene), the correct or full solution can be unequivocally inferred from 100 state transition pairs. Note that for n=50 (genes) and k=3 (inputs per gene) only a small fraction (~100) of all possible state transitions (2^{50} ~ 10^{15} !) is required for reliable inference of the network wiring and rules.

Input entropies

H(A)	1.00
H(B)	1.00
H(C)	1.00
H(A,B)	2.00
H(B,C)	2.00
H(A,C)	2.00
H(A,B,C)	3.00

$$H(X) = - \sum p(x) \log p(x)$$
$$H(X,Y) = - \sum p(x,y) \log p(x,y)$$

$$M(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$M(X,[Y,Z]) = H(X) + H(Y,Z) - H(X,Y,Z)$$

Determination of inputs for element A ①

H(A')	1.00				
H(A',A)	2.00	M(A',A)	0.00	M(A',A) / H(A')	0.00
H(A',B)	1.00	M(A',B)	1.00	**M(A',B) / H(A')**	**1.00**
H(A',C)	2.00	M(A',C)	0.00	M(A',C) / H(A')	0.00

② **Rule table for A**
rule no. 2

input	output
B	**A'**
0	0
1	1

Determination of inputs for element B ③

H(B')	0.81				
H(B',A)	1.50	M(B',A)	0.31	M(B',A) / H(B')	0.38
H(B',B)	1.81	M(B',B)	0.00	M(B',B) / H(B')	0.00
H(B',C)	1.50	M(B',C)	0.31	M(B',C) / H(B')	0.38
H(B',[A,B])	2.50	M(B',[A,B])	0.31	M(B',[A,B]) / H(B')	0.38
H(B',[B,C])	2.50	M(B',[B,C])	0.31	M(B',[B,C]) / H(B')	0.38
H(B',[A,C])	2.00	M(B',[A,C])	0.81	**M(B',[A,C]) / H(B')**	**1.00**

④ **Rule table for B**
rule no. 14

input		output
A	**C**	**B'**
0	0	0
0	1	1
1	0	1
1	1	1

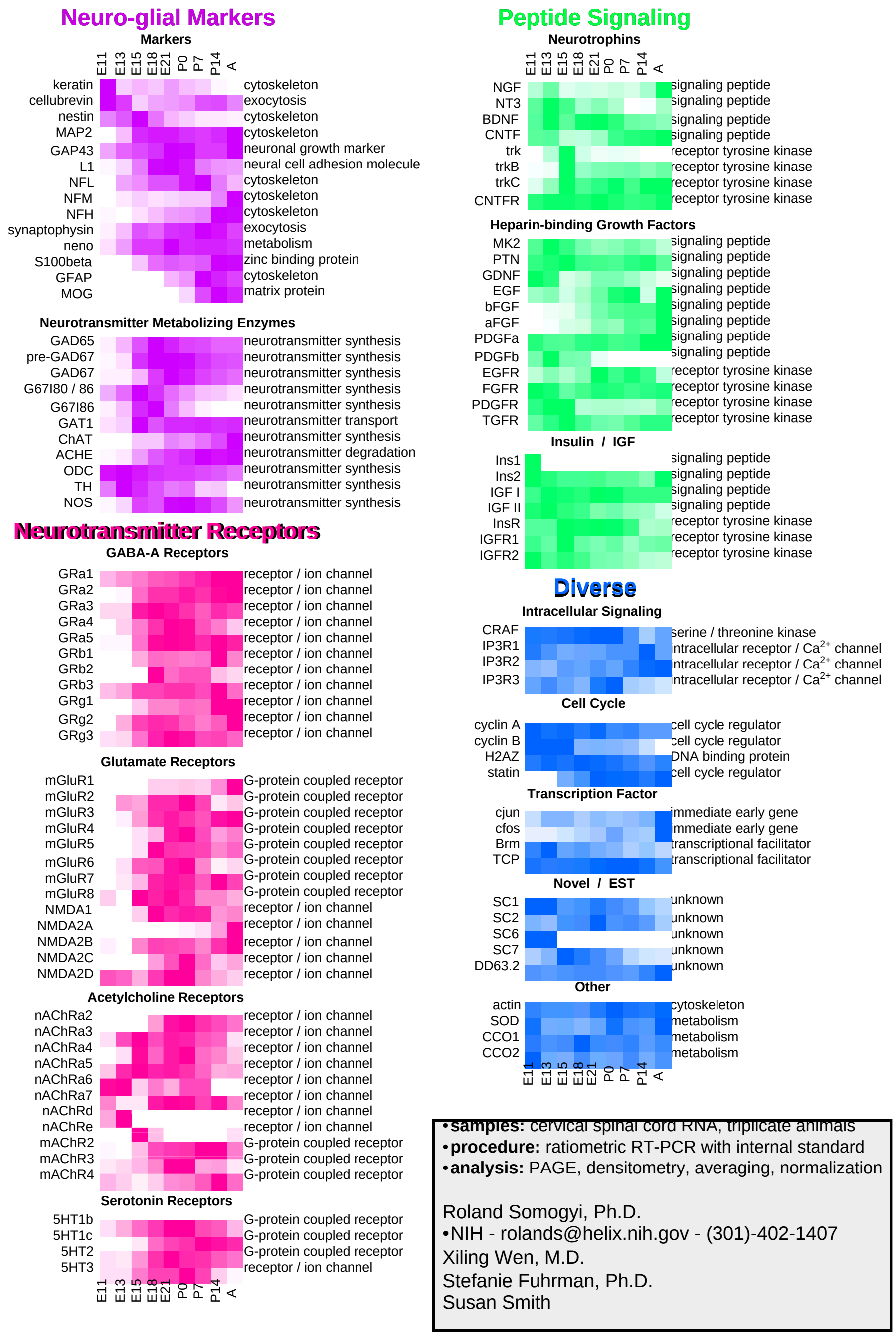
Determination of inputs for element C ⑤

H(C')	1.00				
H(C',A)	1.81	M(C',A)	0.19	M(C',A) / H(C')	0.19
H(C',B)	1.81	M(C',B)	0.19	M(C',B) / H(C')	0.19
H(C',C)	1.81	M(C',C)	0.19	M(C',C) / H(C')	0.19
H(C',[A,B])	2.50	M(C',[A,B])	0.50	M(C',[A,B]) / H(C')	0.50
H(C',[B,C])	2.50	M(C',[B,C])	0.50	M(C',[B,C]) / H(C')	0.50
H(C',[A,C])	2.50	M(C',[A,C])	0.50	M(C',[A,C]) / H(C')	0.50
H(C',[A,B,C])	3.00	M(C',[A,B,C])	1.00	**M(C',[A,B,C]) / H(C')**	**1.00**

⑥ **Rule table for C**
rule no. 170

input			output
A	**B**	**C**	**C'**
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Figure 10 REVEAL principles. Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.



- **samples:** cervical spinal cord RNA, triplicate animals
- **procedure:** ratiometric RT-PCR with internal standard
- **analysis:** PAGE, densitometry, averaging, normalization

Roland Somogyi, Ph.D.
•NIH - rolands@helix.nih.gov - (301)-402-1407
Xiling Wen, M.D.
Stefanie Fuhrman, Ph.D.
Susan Smith

Figure 11**.** Temporal expression patterns for 112 genes expressed in rat spinal cord, as determined by RT-PCR. (Wen et al. (1998). Proc Natl Acad Sci USA, 95:334-339.)

4. Fundamental gene expression data visualization and analysis

4.1. Visualization of gene expression patterns using the density plot

By measuring relative gene expression levels for multiple genes at multiple time points, we obtain a measure of the dynamic output of a genetic network. The time points for Fig. 11 were selected because of their appropriateness to the time scale of rat central nervous system development. Time points range from embryonic day 11 (E11, when spinal cord development begins) to adult or postnatal day 90(P90). Darkest color, highest expression level detected; white, undetectable. For each gene, expression levels are normalized to maximal expression for that gene. Each data point (colored square) is the average of results from three animals. Although the genes we selected are only a tiny fraction of the total number of genes expressed during spinal cord development, they are representative of the different types of proteins that are expressed during the differentiation of tissue. The neurotransmitter- and peptide-related genes encode the proteins and peptides responsible for intercellular communication during CNS development. We have focused on intercellular signaling genes, since these are directly responsible for differentiation in a multicellular organism.

4.2. Cluster and visualization using the dendrogram

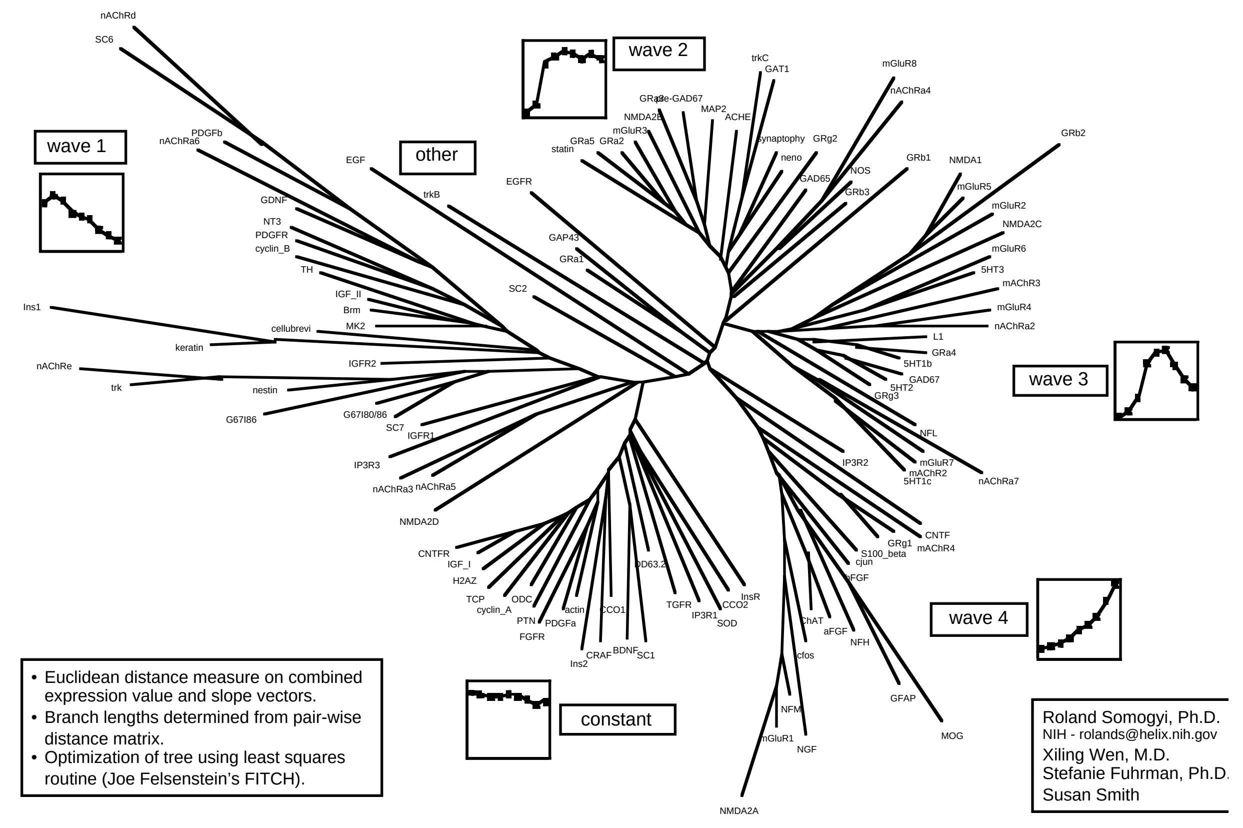
Genes with similar temporal expression patterns may share common genetic control processes, and may therefore be related functionally. Clustering gene expression patterns according to a heuristic distance measure is the first step toward constructing a "wiring diagram" for a genetic network. Such diagrams should permit the development of new hypotheses concerning gene interactions during recovery from injury and disease, or in therapeutic drug treatments.

Exhaustive reverse engineering--the inference of wiring diagrams and functional rules from expression data--is only possible for model networks with current algorithms and limited data sets. However, careful experimental designs and optimization of inference tools may allow significant progress in the future.

4.2.1. Euclidean cluster analysis

- Euclidean distance: A gene expression pattern over n time points is a point in n-dimensional parameter space:
  Distance = $\sqrt{}\Sigma(a_i-b_i)^2$
- Euclidean cluster analysis implies shared wiring and rules (see above, Fig. 6).

We have used the Euclidean distance measure to group genes according to similarities in their temporal expression patterns. This method is similar to clustering according to positive linear correlations. Figure 12 shows a tree generated by the FITCH clustering algorithm using the Euclidean distance matrix for 112 genes expressed in the rat spinal cord. Similarities in temporal expression patterns are indicated by common branch points. According to visual inspection of the tree, the genes appear to cluster into groups, or "waves." Each wave (shown as an inset) corresponds to an average pattern for all the genes of the corresponding cluster: wave 1 genes are expressed at a high level early in development and then decrease in expression toward adult; wave 2 genes are expressed at low levels early, and then plateau; etc. One cluster, "constant," contains the genes whose expression levels are relatively invariant over the time course. Within each wave, the genes may be said to share the same expression kinetics over the time course E11 (embryonic day 11) to adult (P90 or postnatal day 90). The generation of Euclidean distance trees may



- Euclidean distance measure on combined expression value and slope vectors.
- Branch lengths determined from pair-wise distance matrix.
- Optimization of tree using least squares routine (Joe Felsenstein's FITCH).

Roland Somogyi, Ph.D.
NIH - rolands@helix.nih.gov
Xiling Wen, M.D.
Stefanie Fuhrman, Ph.D.
Susan Smith

Fig. 12. Euclidean distance tree for genes expressed in rat spinal cord. Carr DB, Somogyi R, Michaels G (1997) Statistical Computing and Graphics Newsletter 8(1):20-29.

provide clues as to which genes share a common genetic control process. For example, the members of wave 3 may all be regulated by a particular gene. Clustering trees are a first step toward the generation of a "wiring diagram" for a genetic network.

4.2.2. Mutual information cluster analysis

Mutual information: Most general measure of correlation M(A, B) = H(A) + H(B) - H(A, B). M is the information shared by (temporal gene expression patterns) A and B. H refers to the Shannon entropy ($H = -\Sigma p_i \log p_i$), which for our purposes is a measure of the number of expression levels exhibited by a gene. H is high if a gene shows a large number of expression levels over a time course; H is low if the expression pattern is relatively invariant over time. The higher the value of H, the more information the pattern contains. A completely flat, or constant, expression pattern carries no information, and has an H of zero. Note: H reveals nothing about specific expression levels at individual time points, because it is based only on the relative frequencies, i.e. the probability of occurrence, (p_i), of expression levels within a time course.

"Coherence" (normalized mutual information): Captures similarities in patterns independent of individual information entropies. "In how far is pattern A able to predict pattern B?" C = M(A, B) / Hmax(A, B). Coherence is an important consideration because mutual information increases with entropy. We can correct for this bias by dividing the mutual information by the maximum entropy of the pair. For example, even though two genes with relatively "constant" expression patterns have low Hs (and therefore, low M) we may wish to use C to acknowledge that they nevertheless have highly similar patterns.

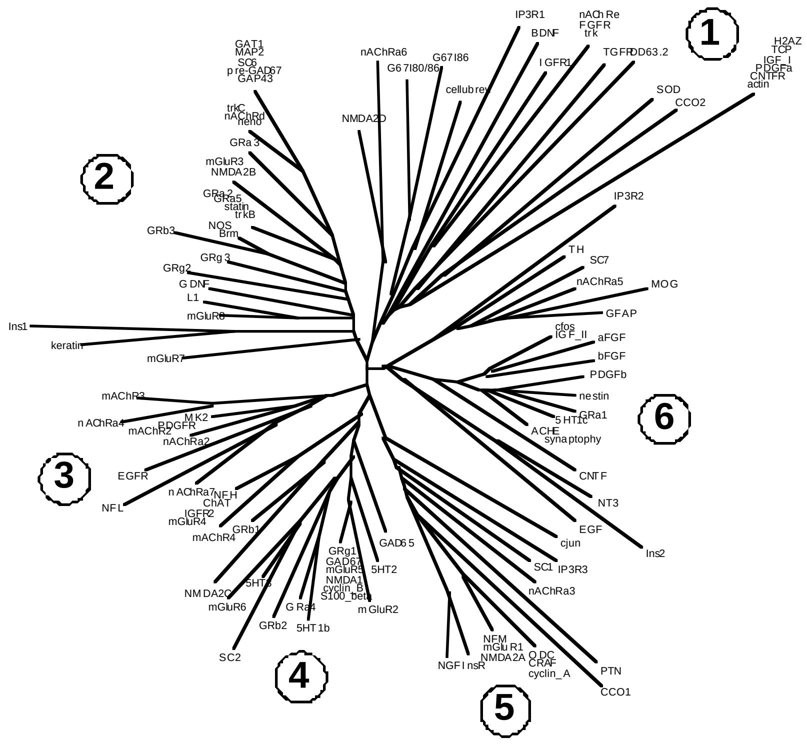


Fig. 13. Mutual information tree for genes expressed in rat spinal cord. Michaels G, Carr DB, Wen X, Fuhrman S, Askenazi M, Somogyi R (1998) Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. Pacific Symposium on Biocomputing 3:42-53.

Mutual information (Coherence) cluster analysis implies shared wiring, with no constraints on rules (Fig 13). Unlike the Euclidean distance measure, mutual information determines negative and nonlinear, as well as positive and linear, correlations. Mutual information therefore clusters genes that may share inputs, but respond to those inputs with different kinetics. For example, genes A and B may receive an input from C (shared wiring), but A would respond to C by increasing, while B would respond by decreasing (different rules). This relationship between A and B would be recognized by mutual information, but not by the Euclidean distance measure.

4.3. Visualization of connectivity across functional categories

Developmental gene expression exhibits apparent redundancy, i.e. is far from maximally diverse. The fact that we have been able to cluster more than one hundred genes into a small number of temporal expression patterns suggests that the number of genetic control processes is much smaller than the number of regulated genes. Of course, the present sample of genes is quite small compared with the whole genome (estimated at 60,000 genes in the rat). Without further studies, it is impossible to know whether any of the remaining genes exhibit as-yet unobserved expression patterns, such as a U-shaped pattern.

4.3.1. Correspondence between expression clusters and general functional classes

It is interesting to note that two Euclidean distance clusters (waves) consist almost entirely of neurotransmitter signaling genes from both the ionotropic and metabotropic classes (Fig. 14). This particular category of genes therefore appears to be confined to specific genetic control processes, and may share a functional role in development

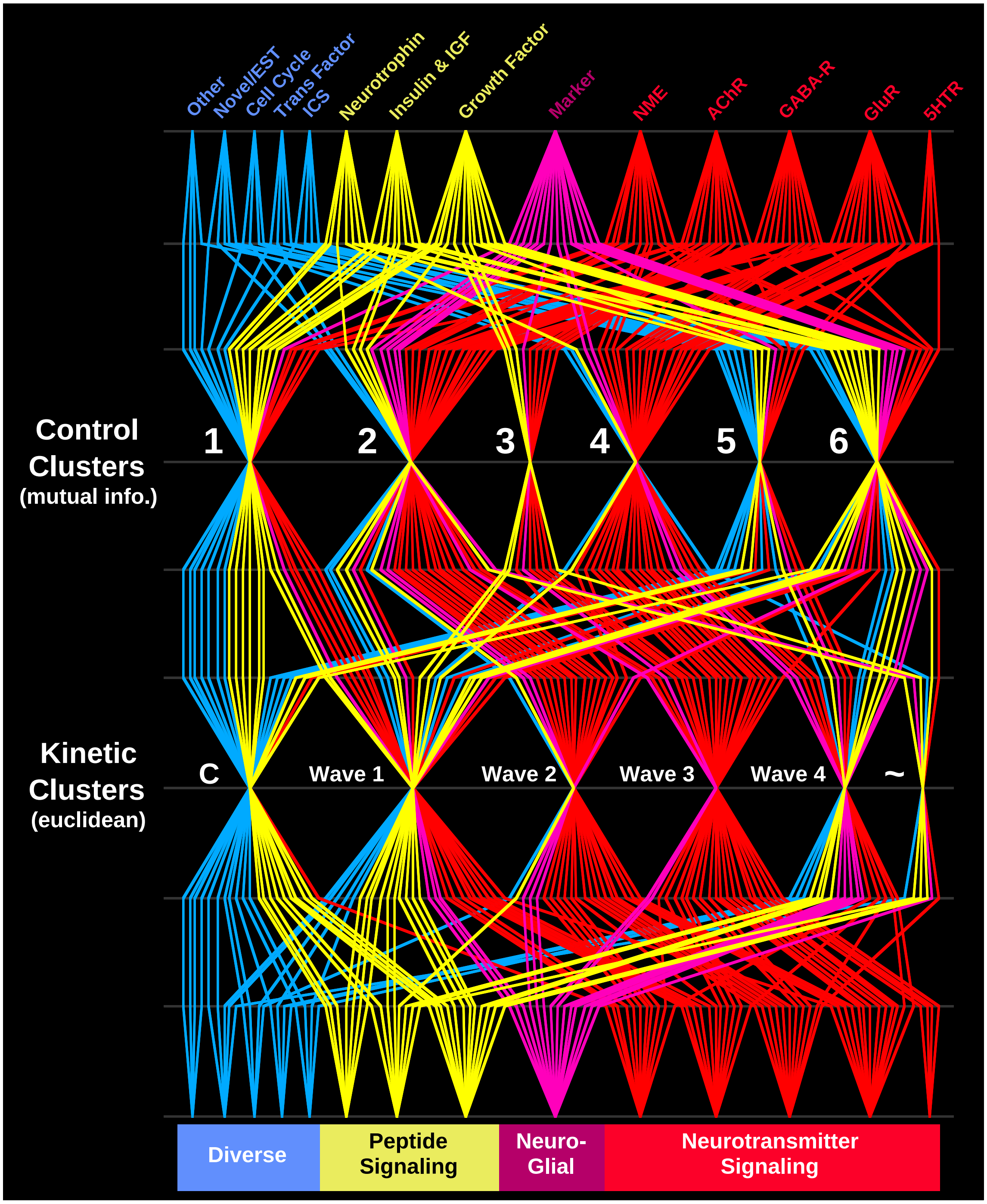


Fig. 14 Functional gene families map to distinct control processes. Michaels G, Carr DB, Wen X, Fuhrman S, Askenazi M, Somogyi R (1998) Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. Pacific Symposium on Biocomputing 3:42-53.

despite differences in DNA sequences. It is also interesting that while another cluster, Euclidean wave 1, contains neurotransmitter signaling genes, these are exclusively ionotropic, suggesting that some ionotropic receptors have a function distinct from that of other neurotransmitter receptors (Fig. 15).

Some proportion of the 60,000 genes of the rat will fall into the constant category, having relatively invariant expression levels over time. For our purposes, constant genes convey no information about phenotypic change, although they may be necessary for maintenance. In higher organisms, it will be necessary to focus on intercellular signaling genes, as these are essential for development in multicellular organisms. Some genes (involved in both intra- and intercellular signaling) can be expected to fall into the constant cluster, even after a perturbation, and may subsequently be ignored. This will allow researchers to concentrate their efforts on the genes most relevant to development or other phenotypic change, such as the response to a disease, injury, or therapeutic drug treatment.

4.3.2. Correspondence between expression clusters and neurotransmitter receptor gene families

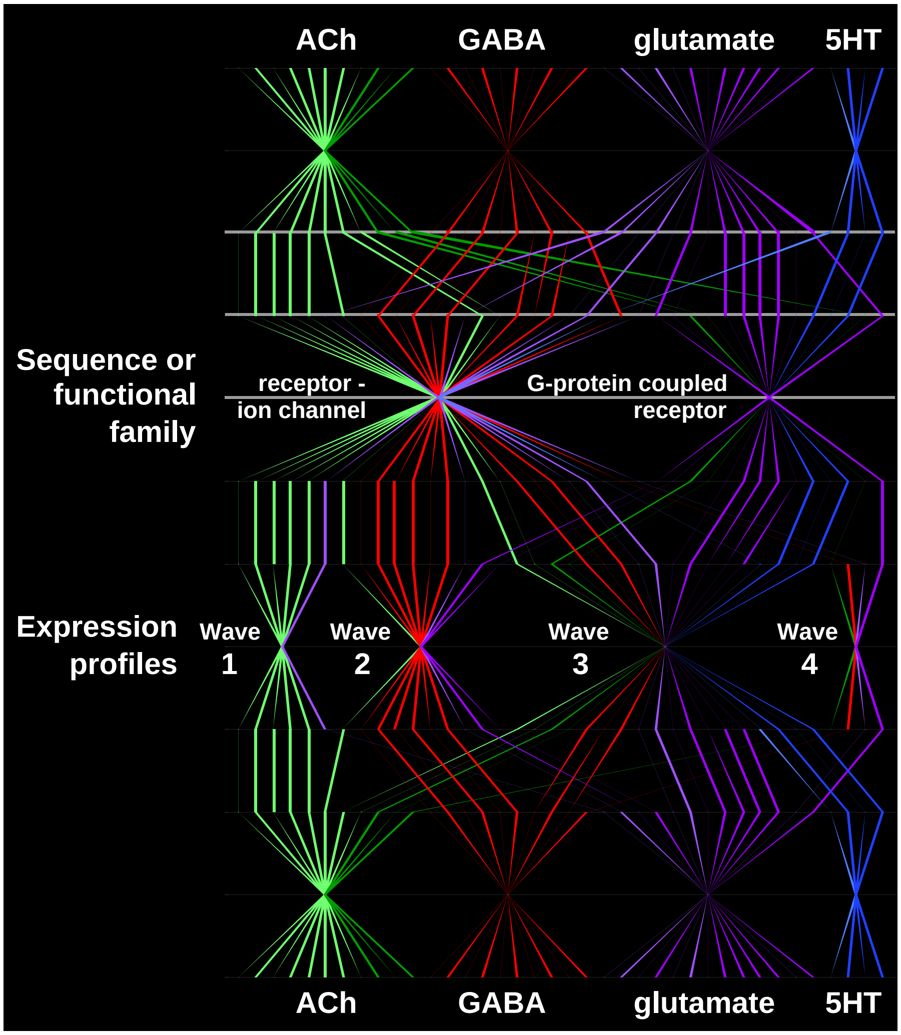


Fig. 15. Neurotransmitter receptors follow particular expression waveforms according to ligand and functional class. See Euclidean distance tree (above) for pictograms showing typical expression profile for each wave. Note that the early expression waves 1 and 2 are dominated by ACh and GABA receptors, and by receptor ion-channels in general. Agnew, B (1998) Science, 280:1516-1518

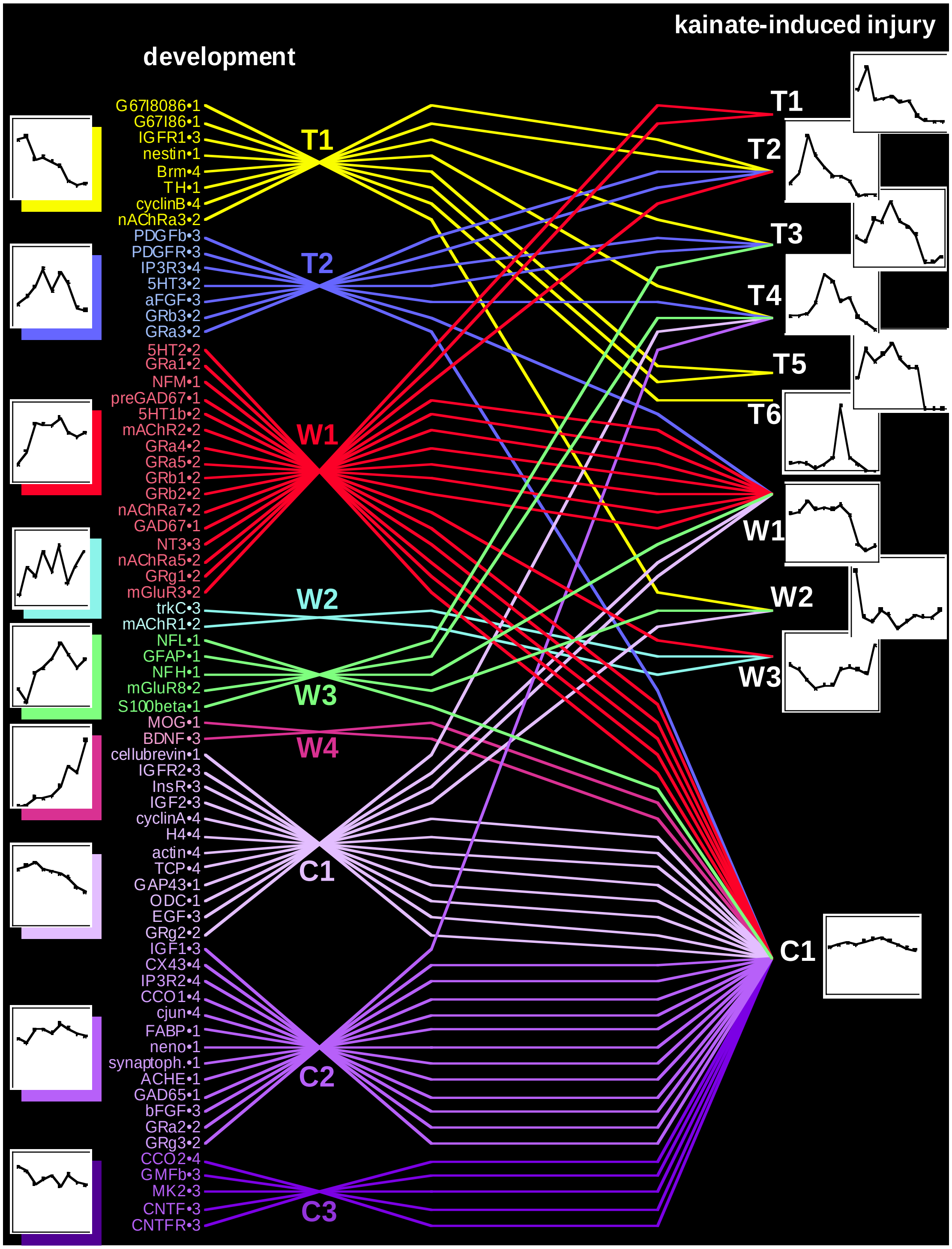


Fig. 16. Mapping of hippocampal developmental gene expression clusters to KA-injury clusters

4.3.3. Mapping of developmental expression patterns to injury-induced expression responses: recapitulation of developmental programs

In Fig. 16, average expression patterns for all clusters are shown as pictograms. Colors correspond to developmental clusters (matches shadowing of developmental pictograms). Lines connect genes in developmental clusters to their respective KA-injury clusters. Each gene can be followed from its label (left column) along a line connecting it to the first focus (developmental cluster) and then, according to the mirror image of this line, to the focus of the KA-injury cluster. Clusters are labeled by Ts, Ws and Cs, corresponding to "Transient," "Waveform," and "Constant" patterns. In development, Ts mark genes that are expressed at significantly higher levels during early to mid development in relation to adult; Ws indicate genes that show other fluctuating patterns; and Cs mark clusters that exhibit relatively invariant expression over the time course. Note that T, W and C cluster members in development generally map to the corresponding T, W and C patterns following KA-injury. We could describe this as a recapitulation of developmental programs in response to a perturbation (seizure). This result suggests that genes may operate within expression *modules*, and provides a clue about the organization of the genetic network.

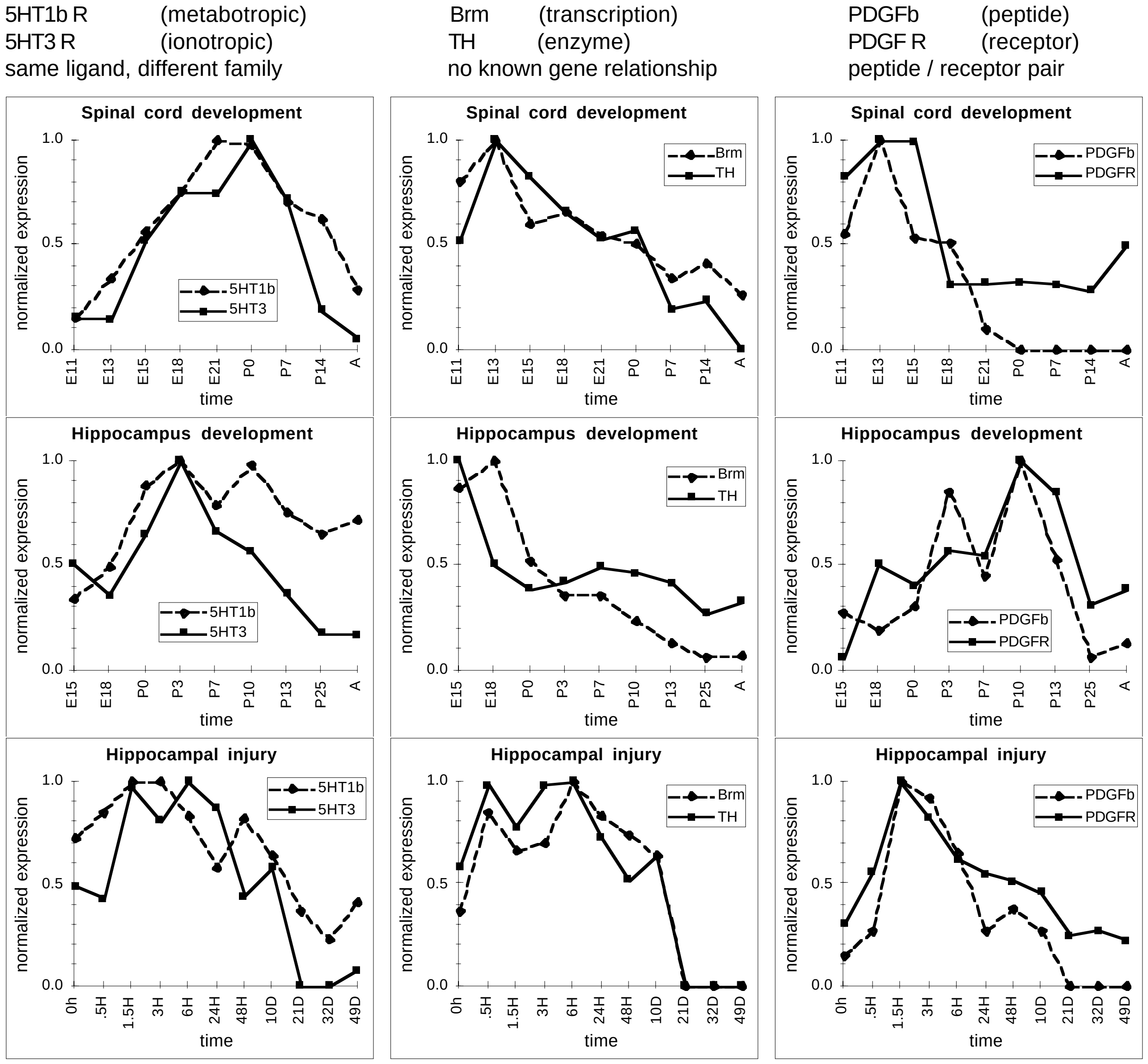


Fig. 17. Gene co-expression pairs in CNS development and injury.

4.3.3. Patterns of global gene co-regulation in rat CNS

Gene expression time series may be useful in determining putative functional connections between genes. As shown in the above figure 17, we found that some pairs of genes exhibit parallel expression patterns under three different conditions: spinal cord development (top), hippocampal development (middle), and hippocampal injury (bottom). Such parallelism may not be surprising in the case of, for example, PDGFb and its receptor. However, to our knowledge, there is no known functional relationship between the transcription factor Brm and the neurotransmitter metabolizing enzyme TH. It is particularly interesting that Brm and TH continue to fluctuate in parallel even after an injury perturbation (chemically-induced seizure). Further studies will be necessary to confirm a functional connection between these two genes. Gene expression time series may also be useful in establishing possible functions for newly discovered genes. This is particularly relevant now that whole genomes are being sequenced. A large proportion of yeast (S. cerevisiae) genes, for example, have no known homologues in other organisms, leaving molecular biologists clueless as to their functions. Large-scale temporal gene expression studies in different tissues and under different conditions can provide a starting point for investigations of these novel genes by comparing their expression time series with those of known genes.

5. Continuous expression data modeling and reverse engineering

5.1. Advantages of continuous models over Boolean

Examining some of the publicly available gene expression data sets, it is clear that genes spend a lot of their time at intermediate values: gene expression levels tend to be continuous rather than discrete, and discretization can lead to a large loss of information. Furthermore, important concepts in control theory that seem indispensable for gene regulation systems either cannot be implemented with discrete variables, or lead to a radically different dynamical behavior: amplification, subtraction and addition of signals; maintaining equilibrium using negative feedback; smoothly varying an internal parameter to compensate for a continuously varying environmental parameter; smoothly varying the period of a periodic phenomenon like the cell cycle, etc.

Granted, some of these problems can be alleviated by hybrid Boolean systems. In particular, Glass has proposed sets of piecewise linear differential equations, where each gene has a continuous-valued internal state, and a Boolean external state. Thomas has proposed an asynchronously updated logic with intermediate threshold values. These systems allow for easy analysis of certain properties of networks, but still do not seem appropriate for quantitative modeling of real gene networks.

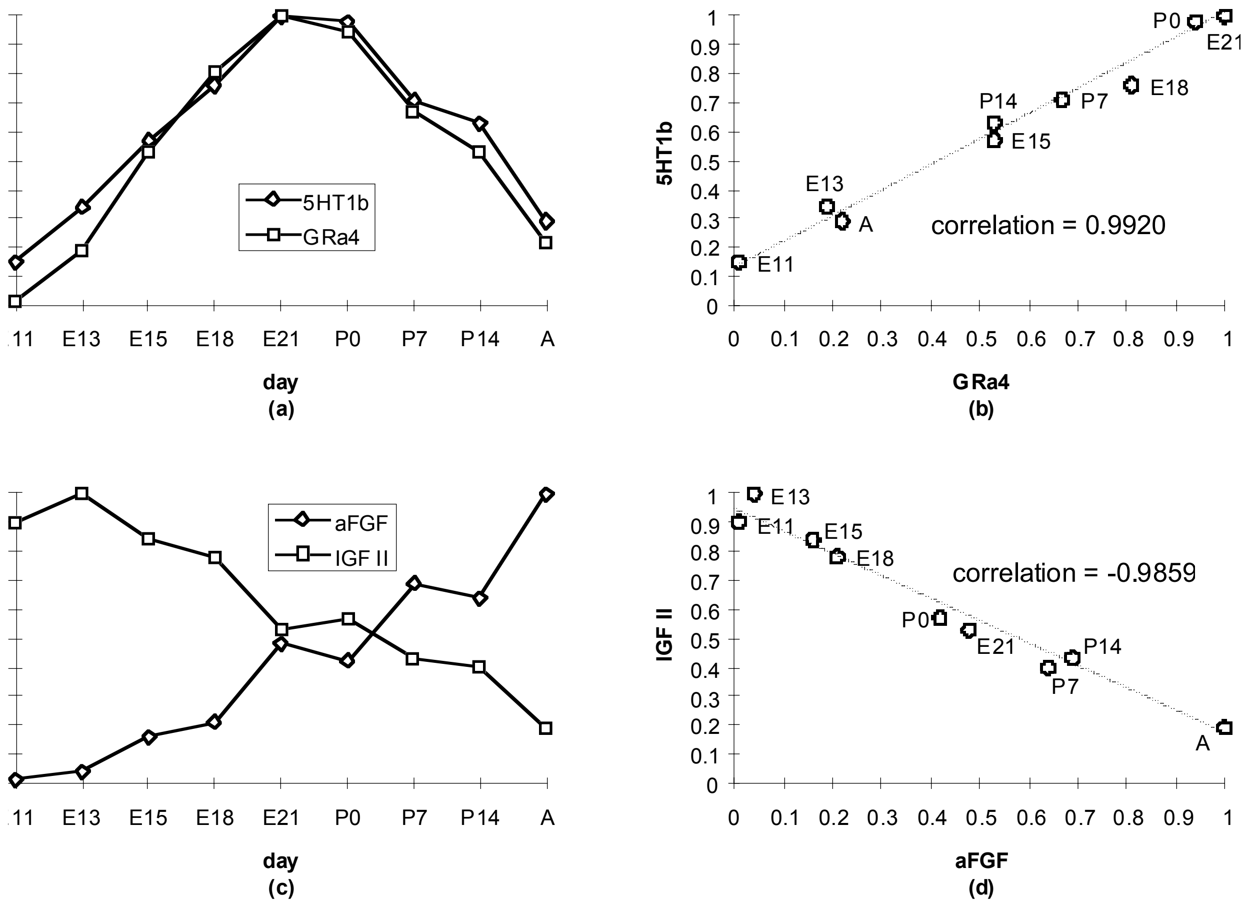


Fig. 18. High positive correlation between gene expression patterns 5HT1b and GRa4 (a, b); high negative correlation between aFGF and IGF II (c, d). Expression levels for each gene are normalized with respect to the maximum expression level for that gene. D'haeseleer P, Wen X, Fuhrman S, Somogyi R (1998) Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data. Proceedings of the International Workshop on Information Processing in Cells and Tissues 1997, pp. 203-212 , Plenum Press..

5.2. Correlation analysis

5.2.1. Linear and rank correlation

Observation of correlations between variables has long been used in biology to predict causal relationships. Although correlation can never provide proof of a causal relationship, it can lead us to propose hypotheses that can be tested by other means. In terms of gene regulation, a high correlation (or anti-correlation) between A and B can be caused by (1) gene A regulating gene B, (2) gene B regulating gene A, (3) gene A and B being co-regulated by a third gene C, or (4) accident. Of course, all of these regulatory interactions can be indirect, through one or more intermediates. Nevertheless, a sufficiently high correlation between two genes (taking into account number of data points, error levels on the data, general regulation trends, etc.) warrants an investigation of the genes in question.

We have presented a preliminary statistical analysis of the rat spinal cord data set, in which relationships between individual genes were inferred based on both linear correlation and rank correlation. Rank correlation allows the detection of tight but nonlinear relationships between variables. Several gene pairs with very high linear correlation were identified (from 0.992 to -0.986), as well as a number of genes with high rank correlation but small linear correlation.

Positive linear correlation is related to Euclidean distance between expression patterns, but is not sensitive to an absolute shift in expression levels. Negative correlation, even though it may indicate a strong linkage between genes, will not show up in a Euclidean distance analysis at all. Clustering based on correlation (using the residual variance, $1-r^2$) may give a better appreciation of co-regulation among genes.

5.2.2. Correlation Metric Construction

Adam Arkin and John Ross at Stanford University have been working on a method called Correlation Metric Construction, to reconstruct reaction networks from measured time series of the component chemicals This approach is based in part on electronic circuit theory, general systems theory and multivariate statistics.

The system (a reactor vessel with chemicals implementing glycolysis) is driven using random (and independent) inputs for some of the chemical species, while the concentration of all the species is monitored over time. First, the time-lagged-correlation (cross correlation) matrix is calculated, and from this a distance matrix is constructed based on the maximum correlation between any two chemical species. This distance matrix is then fed into a simple clustering algorithm to generate a tree of connections between the species. To visualize the results, the chemical species and the tree connecting them is displayed using multidimensional scaling (MDS), mapping each species to a point in 2D space while trying to preserve the distances between each prescribed in the distance matrix. It is also possible to use the information regarding the time lag between species at which the highest correlation was found, which could be useful to infer causal relationships. More sophisticated methods from general systems theory, based on mutual information, could be used to infer dependency.

5.3. Reverse engineering: network inference

5.3.1. Linear and quasi-linear model

The correlation analysis from section 5.2.1. can trivially be extended to finding the subset of genes whose weighted sum correlates best with the expression levels of a specific gene of interest:

$$x_i(t+1) = \Sigma w_{ji}\, x_j(t) + b_i$$

Where x_i is the expression level of gene i at time t, b_i is a bias term indicating whether gene x_I is expressed or not in the absence of regulatory inputs, and weight w_{ji} indicates the influence of gene j on the regulation of gene i. We can rewrite this as a difference equation:

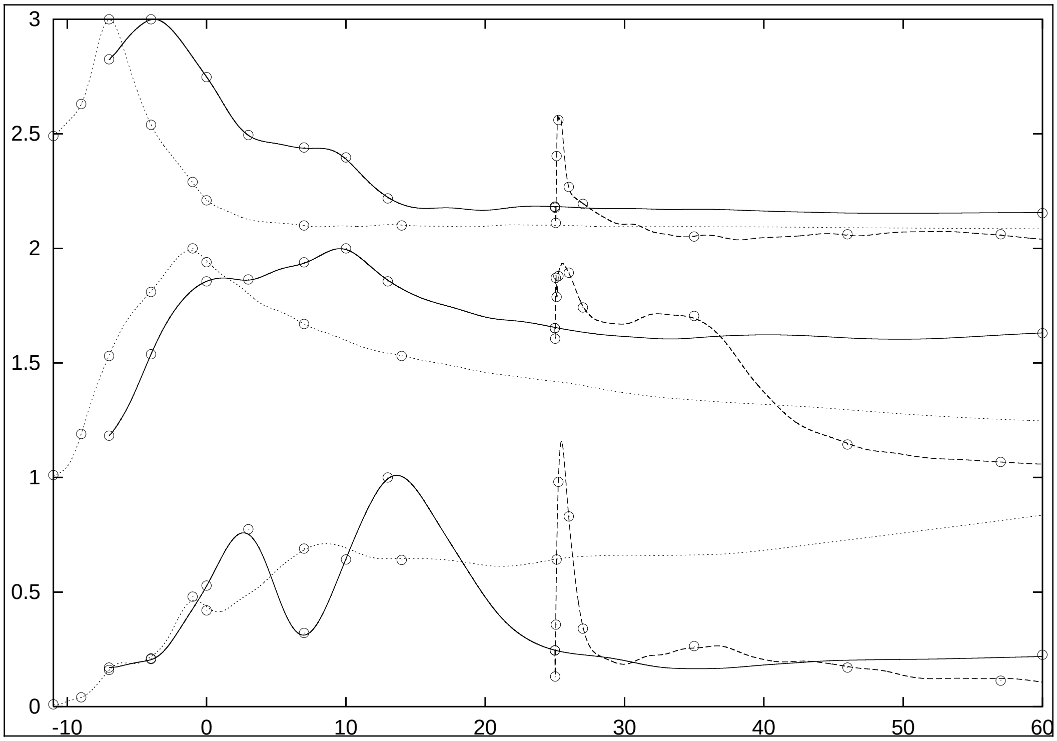


Fig. 19. Reconstructed gene expression time series for nestin (top), Gra4 (middle) and aFGF (bottom). Nestin and Gra4 levels are offset by 2.0 and 1.0 respectively. Time is in days from birth (embryonic day 22). Dotted line: spinal cord, starting at embryonic day 11. Solid line: hippocampus development, starting at embryonic dat 15. Dashed line: hippocampus kainate injury, starting at postnatal day 25. Circles: original data points. P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi (1999) Linear Modeling of mRNA Expression Levels During CNS Development and Injury. Pacific Symposium on Biocomputing.

$\Delta x_i = \Sigma w'_{ji}\, x_j + b_i$

Where $w'_{jj} = w_{jj} - 1$. Given a set of expression patterns equidistant in time, we can use linear algebra to solve for the weights w_{ji} to match the data (provided we have more data points than variables).

We have used this approach to model the 65 genes that make up the intersection of the rat spinal cord and hippocampal data set. An extra input to the weighted sum was added to cover differences in gene regulation among the two tissue types, and another to model the effect of kainate on the hippocampus (half of the hippocampal data set consists of a perturbation experiment tracking the changes in gene expression after kainate injection). Combining the two data sets gives us 22 non-uniformly spaced data points. Very finely spaced equidistant data points were derived through interpolation. The interpolation imposes a smoothness constraint on the expression levels between the original data points, so it does buy us some extra information. Of course we would have preferred to have more than 22 data points to model 65 genes in the first place.

Despite the fact that this technique was only borderline feasible due to the limited amount of data, some of the results were quite interesting: The resulting weight matrix turned out to be sparse, in agreement with our intuition that genes are not regulated by all other genes (note that this model is unconstrained with respect to number of

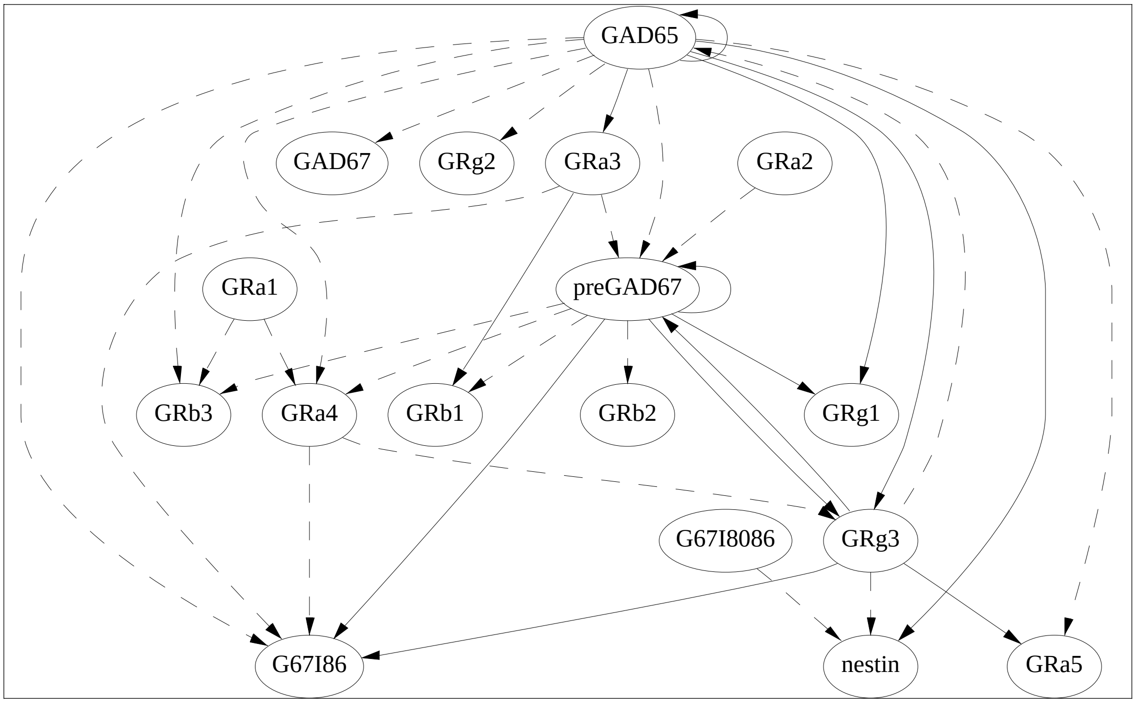


Fig. 20. Subgraph with main interactions between GAD and GABA-receptors, derived from the linear model. P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi (1999) Linear Modeling of mRNA Expression Levels During CNS Development and Injury. Pacific Symposium on Biocomputing.

regulatory interactions); some biologically important genes were regulating many other genes, whereas many others had very few regulatory outputs (the number of regulatory inputs to genes had a much narrower distribution); some of the genes seemed to have a primarily positive or negative regulatory role. Starting with the initial gene expression levels, the model could accurately regenerate the entire trajectories (spinal cord development, hippocampus development and hippocampus injury) by iterating the difference equation for each time step. Eigenvector analysis showed three attractors of the system at the adult spinal cord, adult hippocampus and injured hippocampus expression patterns. However, most of the specific predictions of the model could not be verified because so little is known about regulatory interactions in mammal CNS.

For added biological realism, we can include a sigmoidal squashing function into the equation above:

$x_i(t+1) = g(\Sigma w_{ji}\, x_j(t) + b_l)$

Weaver et al show in a paper in this Symposium that this sort of quasi-linear model can be solved by linear algebra as well, by first applying the inverse of the squashing function:

$g^{-1}(x_i(t+1)) = \Sigma w_{ji}\, x_j(t) + b_l$

They also showed that randomly generated networks can be accurately reconstructed using this modeling technique.

Mjolsness, Reinitz and Sharp have used a similar approach to model small gene networks involved in pattern formation during the blastoderm stage of development in Drosophila. They added a simplified cellular model, with synchronized cell divisions (cell divisions are under the control of a maternal clock at this stage) along a longitudinal axis, alternated with updating the gene expression levels. Because of the more complex hybrid model, simulated annealing was used to find a least-squares fit to real gene expression data. The model was able to successfully replicate the pattern of eve stripes in Drosophila, as well as some mutant patterns on which the model was not explicitly trained.

Various people have coined different names for this sort of models: connectionist model (Mjolsness, Reinitz and Sharp), linear model (D'haeseleer), linear transciption model (Chen et al), weight matrix model (Weaver et al). Considering the core of these models contain a weighted sum to implement gene regulation, perhaps we should call them *additive models*.

5.3.2. Differential equation models

The difference equations from the previous section should remind us that differential equations have been used for years to model known biomolecular interactions on individual operators.

One implicit assumption is that the concentrations of the chemical species are continuous, i.e. that stochastic fluctuations due to single molecules can be ignored. We know that this does not hold at least for some proteins which are present in concentrations of only a couple of molecules per cell. Indeed, there are indications that stochastic fluctuations may actually be exploited by some organisms. However, differential equations are widely used to model biochemical systems. Hopefully, a continuous approach will prove to be appropriate for the majority of interesting mechanisms.

Chen et al at this Symposium present a number of linear differential equation models, including both mRNA and protein levels. They show theoretically how to solve for the parameters of the models using linear algebra (as in 5.3.1.) and Fourier transforms. They find that their model can not be solved from mRNA concentrations alone, without at least the initial protein levels. Conversely, the model can be solved given only a time series of protein concentrations.

5.3.3. Recurrent neural networks

Neural networks have the undeserved reputation of being no more than a black-box modeling tool, without any power to illuminate the underlying structure of a system. However, keep in mind that there exist many entirely different kinds of neural networks. The one that concerns us here is essentially equivalent to a set of differential equations similar to the difference equation listed in 5.3.1.:

$$dx_i/dt = g(\Sigma w_{ji}\, x_j(t) + b_l) - D_i\, x_i$$

where D_i is a decay constant. This is a very accepted, although simplified, differential equation model for gene regulation. What the neural network viewpoint buys us is a way to adjust ("train") the parameters of this set of nonlinear differential equations to match the expression data. Training algorithms exist for both time series data (training on a trajectory) and single-point expression patterns (training on an attractor). Additionally, the training algorithms allow us to incorporate various kinds of a priori information regarding expected degree of connectivity, distribution of connection weights, known or hypothesized interactions, etc. Given proper constraints, it should be feasible to arrive at a neural network model of which the wiring reflects the biological system, given limited data.

4. Bibliography

Publications by the authors

- Carr DB, Somogyi R, Michaels G (1997) Templates for Looking at Gene Expression Clustering. Statistical Computing and Graphics Newsletter 8(1):20-29.
- D'haeseleer P, Wen X, Fuhrman S, Somogyi R (1998) Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data. *Proceedings of the International Workshop on Information Processing in Cells and Tissues* 1997, pp. 203-212 , Plenum Press.
- D'haeseleer. P., X. Wen, S. Fuhrman, and R. Somogyi (1999) Linear Modeling of mRNA Expression Levels During CNS Development and Injury. Pacific Symposium on Biocomputing.
- Fuhrman S, Wen X, Michaels G, Somogyi R (1998) Genetic Network Inference. Interjournal. See www.interjournal.org/
- Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.
- Michaels G, Carr DB, Wen X, Fuhrman S, Askenazi M, Somogyi R (1998) Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. Pacific Symposium on Biocomputing 3:42-53.
- Somogyi R, Sniegoski CA (1996) Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. Complexity 1(6):45-63.
- Somogyi R, Fuhrman S, Askenazi M, Wuensche A (1997) The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. Nonlinear Analysis, Proc. of Second World Cong. of Nonlinear Analysts (WCNA96), 30(3):1815-1824.
- Somogyi, R (1998) Many to One Mappings as a Basis for Life. See www.interjournal.org/
- Somogyi R., Fuhrman, S (1998) Distributivity, a General Information Theoretic Network Measure, or Why the Whole is More than the Sum of its Parts. *Proceedings of the International Workshop on Information Processing in Cells and Tissues* 1997, pp. 273-283 , Plenum Press.

- Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998) Large-Scale Temporal Gene Expression Mapping of CNS Development. Proc Natl Acad Sci USA, 95:334-339.

Overview

- Kauffman, S.A. (1993) The Origins of Order, Self-Organization and Selection in Evolution. Oxford University Press.
- Bryant, A. Milosavljevic and R. Somogyi (1998) Gene Expression and Genetic Networks. Pacific Symposium on Biocomputing 3:3-5 (1998).

Data

- Anderson, L., and Seilhamer, J. (1997) A comparison of selected mRNA and protein abundances in human liver. Electrophoresis18(4):533-537.
- Appel, R. D., Sanchez, J. C., Bairoch, A., Golaz, O., Miu, M., Vargas, J. R., and Hochstrasser, D. F. (1993) SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. Electrophoresis 14(11):1232-1238. (http://www.expasy.ch/ch2d/ch2d-top.html and http://www.expasy.ch/ch2d/2d-index.html)
- DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680-686. (http://cmgm.Stanford.EDU/pbrown/explore/)
- VanBogelen, R. A., Abshire, K. Z., Pertsemlidis, A., Clark, R. L., and Neidhardt, F. C. (1996) Gene-protein database of Escherichia coli K-12: Edition 6. In Escherichia coli and Salmonella typhimurium Cellular and Molecular Biology, F. C. Neidhardt, R. I. Curtiss, C. A. Gross, J. L. Ingraham, and M. Riley, Eds., 2nd ed. ASM Press, Washington D.C. (http://pcsf.brcf.med.umich.edu/eco2dbase/)
- Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Jr., D. E. B., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997) Characterization of the yeast transcriptome. Cell 88:243-251. (http://www.sagenet.org/yeast/yeastintro.htm)
- Zhang, L., Zhou, W., Velculescu, V. E., Kerm, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. (1997) Gene expression profiles in normal and cancer cells. Science 276:1286- 1272. (http://welchlink.welch.jhu.edu/~molgen-g/home.htm)
- Wen X., Fuhrman S., Michaels G.S., Carr D.B., Smith S., Barker J.L., Somogyi R. (1998) Large-Scale Temporal Gene Expression Mapping of CNS Development. Proc Natl Acad Sci 95:334-339. (http://rsb.info.nih.gov/mol-physiol/PNAS/GEMtable.html)
- P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell 9:3273-3297. (http://genome-www.stanford.edu/cellcycle/)
- Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2(1):65-73. (http://genomics.stanford.edu/yeast/cellcycle.html)

- Araceli M. Huerta, Heladia Salgado, Denis Thieffry, Julio Collado-Vides (1998) RegulonDB: a database on transcriptional regulation in Escherichia coli. Nucleic Acids Res 26(1):55-9. (http://www.cifn.unam.mx/Computational_Biology/regulondb/)

High throughput methods

- DeRisi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996) Use of a cDNA microarray to analyze gene expression patterns in human cancer. Nature Genetics 14:457-460.
- Fields, S., and Song, O. (1989) A novel genetic system to detect protein protein interactions. Nature 340:245-246.
- Fodor, S. P. A., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P., and Adams, C. L. (1993) Multiplexed biochemical assays with biological chips. Nature 364:555-556.
- Liang, P., and Pardee, A. B. (1992) Differential display of eukaryotic mRNA by means of the polymerase chain reaction. Science 257:967-971.
- Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467-470.
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996) Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. 93:10614-10619.
- Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. Science 270(5235):484-487.

Asynchronous Boolean networks

- Thieffry, D. and Thomas, R. (1998) Qualitative Analysis of Gene Networks. Pacific Symposium on Biocomputing 3:66-76.
- Thomas, R. (1991) Regulatory Networks Seen as Asynchronous Automata: A Logical Description. J Theor Biol. 153: 1-23.
- Snoussi, E. H., and Thomas, R. (1993) Logical identification of all steady states: the concept of feedback loop characteristic states. Bull. of Math. Biol. 55:973-991.

Continuous logical networks

- Glass, L., and Kauffman, S. A. (1972) Co-operative components, spatial localization and oscillatory cellular dynamics. J. Theor. Biol. 34:219-237.
- Glass, L. and Kauffman, S.A. (1973) The Logical Analysis of Continuous, Non-Linear Biochemical Control Networks. J. Theor. Biol. 39:103-129.
- Glass, L. (1975). Classification of Biological Networks by Their Qualitative Dynamics. J. Theor. Biol. 54:85-107.

Stochastic behavior of networks:

- McAdams, H.H., Arkin, A. (1997) Stochastic Mechanisms in Gene Expression. PNAS, USA 94(3):814.
- Arkin, A., Ross, J., and McAdams, H. H. (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. Genetics 149(4):1633-48.

Biological constraints on genetic feedback networks:

- Barkal, N., and Leibler, S. (1997) Robustness in simple biochemical networks. Nature 387:913-917.
- Hlavacek, W. S., and Savageau, M. A. (1995) Subunit structure of regulator proteins influences the design of gene circuitry: analysis of perfectly coupled and completely uncoupled circuits. J. Mol. Biol. 248(4):739-755.
- Savageau, M. A. (1977) Design of molecular control mechanisms and the demand for gene expression. Proc. Natl. Acad. Sci. 74:5647-5651.
- Savageau, M.A. (1998) Rules for the Evolution of Gene Circuitry. Pacific Symposium on Biocomputing 3:54-65.

Reverse engineering

- Arkin, A., Shen, P., Ross, J. (1997) A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. Science 277:1275-1279.
- Arkin, A., and Ross, J. (1997) Statistical Construction of Chemical Reaction Mechanisms from Measured Time-Series. J. Phys. Chem. 99:970-979.

Bottom-up modeling of small genetic networks

- McAdams, H.H., Shapiro, S. (1995) Circuit Simulation of Genetic Networks. Science. 269:650-656.

Genetic network modeling of spatio-temporal patterns in development

(combined bottom up and top-down approach)

- Mjolsness, E., Sharp, D. H. and Reinitz, J. (1991) A connectionist model of development. J. Theor. Biol. 152: 429-453.
- Reinitz, J., Mjolsness, E., and Sharp, D.H. (1995) Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. J. Exp. Zool. 271: 47-56.
- Reinitz, J. and Sharp, D.H. (1995) Mechanism of eve stripe formation. Mech. Dev.49: 133-158.

Cis-regulatory structures

- Arnone, M.I. and Davidson, E. (1997) The Hardwiring of Development: Organization and Function of Genomic Regulatory Systems. Development 124:1851-1864.

Linear or quasi-linear models

- Mjolsness, E., Sharp, D. H., and Reinitz, J. (1991) A connectionist model of development. J. Theor. Biol. 152(4):429-454.
- T. Chen, H. L. He, and G.M. Church (1999) Modeling Gene Expression with Differential Equations. Pacific Symposium on Biocomputing.
- D.C. Weaver, C.T. Workman, G.D. Stormo (1999) Modeling Regulatory Networks with Weight Matrices. Pacific Symposium on Biocomputing.

General systems theory

- Bonnlander, B. V., and Weigend, A. S. (1994) Selecting input variables using mutual information and nonparametric density estimation. In Proceedings of the 1994 International Symposium on Artificial Neural Networks, 42-50.
- Broekstra, G. (1981) C-analysis of C-structures: representation and evaluation of reconstruction hypotheses by information measures. Int. J. Gen. Sys. 7:33-61.
- Cavallo, R. E., and Klir, G. J. (1981) Reconstructability analysis: overview and bibliography. Int. J. Gen. Sys. 7:1-6.

Neural networks

- Bray, D. (1990) Intracellular signaling as a parallel distributed process. J. Theor. Biol. 143:215-231.
- Mjolsness, E., Sharp, D. H., and Reinitz, J. (1991) A connectionist model of development. J. Theor. Biol. 152(4):429-454.
- Heckerman, D. A (1995) tutorial on learning with bayesian networks. Tech. Rep. MSR-TR-95-06, Microsoft Research, Redmond, WA. Available via ftp from ftp.research.microsoft.com in /pub/Tech-Reports/Winter94- 95/TR-95-06.PS.
- Hofmann, R., and Tresp, V. (1996) Discovering structure in continuous variables using bayesian networks. In Advances in Neural
- Information Processing Systems 8, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds., MIT Press, 500-506.
- Pearlmutter, B. A. (1995) Gradient calculations for dynamic recurrent neural networks: a survey. IEEE Transactions on Neural Networks 6(5):1212-1228.