Pacific Symposium on Biocomputing 4:41-52 (1999)

LINEAR MODELING OF mRNA EXPRESSION LEVELS DURING CNS DEVELOPMENT AND INJURY

P. D'HAESELEER
University of New Mexico, Department of Computer Science,
Albuquerque, NM 87131, USA

X. WEN, S. FUHRMAN, R. SOMOGYI
Incyte Pharmaceuticals, Inc., 3174 Porter Dr.,
Palo Alto, CA 94304, USA

Large-scale gene expression data sets are revolutionizing the field of functional genomics. However, few data analysis techniques fully exploit this entirely new class of data. We present a linear modeling approach that allows one to infer interactions between all the genes included in the data set. The resulting model can be used to generate interesting hypotheses to direct further experiments.

1 Introduction

With the advent of the "Age of Genomics" an entirely new class of data is emerging. As the goal of *structural genomics*—sequencing entire genomes—comes into sight, the focus is gradually shifting to *functional genomics*. One of the important tools in functional genomics is the large-scale gene expression assay. Using advanced molecular biology techniques, it has become possible to measure the gene expression levels (mRNA levels) of most, if not all, of the genes of an organism simultaneously. The driving force behind this data collection effort is the hope that we might be able to reconstruct the underlying gene regulation networks from it. Progress in this field could have deep implications in bioengineering and therapeutic target discovery.

Wen *et al*[1] have published a Gene Expression Matrix of 112 mRNA species measured at nine different stages during the development of rat cervical spinal cord. Recently, the same team developed a similar data set [2] of 70 mRNA species measured at nine time points during development of rat hippocampus, and at ten more time points following injury of the central nervous system by injection with kainate, a glutamatergic agonist which causes seizures, localized cell death, and severely disrupts the normal gene expression patterns. These data sets are currently the largest publicly available gene expression time series in terms of number of time points, using a high fidelity gene expression assay.

Considering the large amount of overlap between the mRNA species for these two data sets (65 species in common) and the related tissue types (rat cervical spinal cord and hippocampus), it is possible to join this data into one

larger data set of 65 genes by 28 time points, consisting of 1) cervical spinal cord development, 2) hippocampus development, and 3) hippocampus injury. After all, the regulatory "hardware" of the genes is the same, though different parts of it might be active in different contexts. Combining data from different tissues allows us to get a more complete picture of the regulatory interactions.

Previous analyses of the data from Wen *et al*[1] mainly looked at similarities between expression patterns based on Euclidean distance [1,3], linear and rank correlation [4] and information theory based measures [4]. Other approaches at inferring gene networks from time series include boolean network models,[5] Correlation Metric Construction,[6] modeling spatial differentiation,[7] and quantification of a known metabolic pathway.[8]

A biological system can be considered to be a state machine, where the change in internal state of the system depends on the current internal state plus any external inputs. The mRNA levels form an important part of the internal state of a cell (ideally, we also want to measure protein levels, metabolites, etc.). As a first approximation, we fit the expression data with a purely linear model, where the change in expression level of each mRNA species is derived as a weighted sum of the expression levels of all other genes. Of course, a linear model can never be much more than a caricature of the real system, but perhaps we can still draw some interesting conclusions from it. The value of a coarse model like this is mainly exploratory. It serves to direct further detailed investigation by suggesting novel hypotheses about the system.

Although the ultimate goal of this approach is to deduce the causal relationships between genes—the "wiring pattern" of the underlying gene regulatory network—not all the interactions between genes discovered by the current model will represent direct causal relationships. At a total of 65 measured mRNA species, there are inevitably important intermediate steps missing in the model. Perhaps more importantly, the model does not enforce any measure of economy of connections. So, whereas the real gene network may include genes A and B regulating gene C, which then regulates genes D and E, the model could have connections from A and B directly to C, D and E, with no sign of regulation by C (simply because such a pattern of connections may allow the model to better fit the given data set).

Of course, the exact mechanism of regulation of each individual gene cannot be elucidated by this approach. Other, more classical methods exist to tease apart the regulation machinery of a single gene. We are more interested in systemic gene regulation aspects: What is the overall pattern of gene regulation, including feedback circuits, signaling cascades, etc.? Which classes of genes regulate or are regulated by which other classes? Which genes regulate or are regulated in similar ways?

2 The Linear Model

The basic linear model is of the form

$$X_i(t + \Delta t) = \sum_j W_{ij} X_j(t) \tag{1}$$

where $X_i(t + \Delta t)$ is the expression level of gene i at time $t + \Delta t$, and W_{ij} indicates how much the level of gene j influences gene i. For each gene, we will also add an extra term indicating the influence of kainate, and a constant bias term to model the activation level of the gene in the absence of any other regulatory inputs. The differences in gene regulation due to tissue type will be modeled by a difference in bias. The final formula becomes:

$$X_i(t + \Delta t) = \sum_j W_{ij} X_j(t) + K_i \cdot kainate(t) + C_i + T_i. \tag{2}$$

where $kainate(t)$ is the kainate level at time t, K_i is the influence of kainate on gene i, C_i is a constant bias factor for each gene, and T_i indicates the difference in bias between tissue types ($T_i = 0$ when simulating spinal cord, so the total bias for spinal cord is C_i, for hippocampus $C_i + T_i$).

This can be rewritten as a difference equation:

$$\frac{X_i(t + \Delta t) - X_i(t)}{\Delta t} = \sum_j T_{ij} X_j(t) + K'_i \cdot kainate(t) + C'_i + T'_i. \tag{3}$$

where $T_{ij} = (W_{ij} - 1)/\Delta t$ if $i = j$, $T_{ij} = W_{ij}/\Delta t$ otherwise; $K'_i = K_i/\Delta t$; $C'_i = C_i/\Delta t$ and $T'_i = T_i/\Delta t$.

Provided the time step Δt is small enough, T_{ij}, K'_i, C'_i and T'_i will be independent of Δt. Given the time series $X_i(t)$, finding these parameters requires solving a least squares system of linear equations, or, equivalently, performing a multiple regression of each gene on all other genes.

Considering the extremely non-uniform spacing of the measurements (half hour interval after kainate injection, more than two months interval before the final adult cervical spinal cord measurement), we first constructed an interpolated time series from the data using a cubic interpolation on the log of the expression levels. (Taking the log before interpolating prevents negative values in the interpolation.) An interpolation rate of 10 time points per hour gives us 5 interpolated points between the two closest measurements, and still allows us to calculate the least squares fit over the entire 7-month data set.

3 Results

Note that the original data set, 65 genes by 28 time points, is really too small to be fit by a linear model with 65×68 parameters (we would need at least 68 time points to do so). Using linear interpolation between the time points, the model would indeed be underconstrained: an infinite number of different linear models would fit the data. However, because the nonlinear interpolation scheme takes into account non-local information, we do arrive at 65 linearly independent time series after interpolation. The smallest singular value (indicating the degree of linear independence) for the total data set is 0.028, more than an order of magnitude better than for both hippocampus data sets, and three orders of magnitude better than for the spinal cord data set or either of the hippocampal data sets by itself (using linear interpolation, all but 28 singular values would be equal to zero). This indicates that a linear model of the combined data sets will be significantly less underconstrained. Still, because of the limited number of original data points, the results obtained here are only speculative, and are intended primarily to illustrate the method.

To evaluate the accuracy of this modeling approach we would want to apply it to a system of which the gene regulation network is already well understood. Unfortunately, comparatively little is known regarding gene regulation of these 65 genes in CNS development, making a direct evaluation infeasible. Part of this section, especially Subsection 3.1, will be devoted to circumstantial evidence showing that the resulting model may indeed be a reasonable representation of the underlying regulation network.

3.1 A Biologically Plausible Model?

The histogram of interaction weights T_{ij} resulting from the least squares fit to the interpolated time series is very sharply peaked around zero (see Fig. 1). This means the connection matrix is a good approximation to a sparse matrix, i.e., each gene is only influenced by a limited number of others, as we would expect for the "real" connection matrix.

There are five genes which have a disproportionately large input vector (i.e. a large number of parameters T_{ij} for gene i are nonzero): BDNF, G67I8086, GFAP, GRa1 and NFM. All these genes have input vector sizes larger than 12, compared to an average input vector size of 4.20. Perhaps these genes are simply regulated by a large number of different factors. More likely, the genes are inadequately modeled with a linear approach, either because their regulation is highly nonlinear, some of their regulating factors are not in the data set, or a variety of other reasons. So far, it is unclear why precisely these five genes should be modeled poorly. When discussing interaction weights in

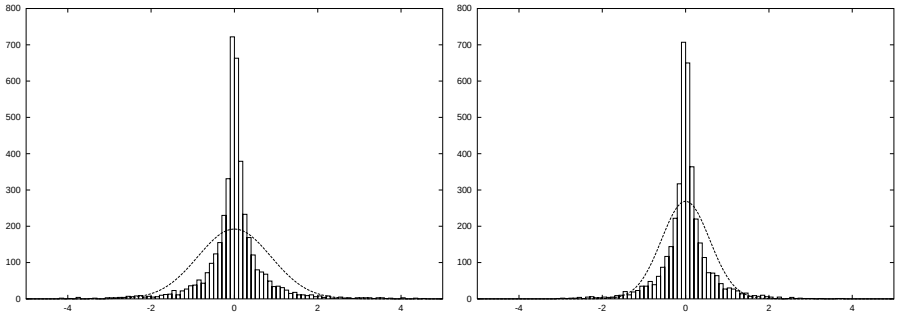

Figure 1: Histogram of interaction weights versus a Gaussian of equal standard deviation. Right: histogram without the input weights for BDNF, G67I8086, GFAP, GRa1 and NFM.

the remainder of this paper, we will leave out the inputs to these genes. Since the least squares solution essentially solves a linear regression for each gene independently, failure to achieve a biologically plausible model for some of the genes does not imply that the rest of the model is unreliable.

The sum of input weights to each gene is close to zero, i.e. there seem to be no genes that are primarily positively or negatively influenced by other genes. However, the sum of output weights from each gene varies more widely, with a standard deviation that is almost a magnitude larger than for the sum of input weights, indicating that some genes have a primarily negative or positive influence. This is in agreement with our biological knowledge, because many genes are known to have a primarily up-regulating or down-regulating role. According to the model, the major regulating genes are mGluR3, AChE, 5HT1b, GRa2, preGAD67 and GAD65. Note that this short list consists entirely of neurotransmitter metabolizing enzymes (AChE, GAD67, GAD65) and neurotransmitter receptors (mGluR3, 5HT1b, GRa2).

mGluR3 is a member of the metabotropic glutamate receptor family, and transduces the glutamate signal to the intracellular signaling biochemistry. It is not known whether it plays a more central role than the other mGluR's, so this may point to an interesting hypothesis. (However, note that at least 6 other mGluR's are missing from the 65 genes in the intersection of the spinal cord and hippocampus data sets, so perhaps mGluR3 is also filling in for some of the missing glutamate receptors.) Furthermore, mGluR3 is a G-protein coupled receptor that inhibits adenylate cyclase, leading to a reduction of cAMP,[10] a general intracellular effector which is involved in multiple signaling pathways. If cAMP is a positive modulator of the genes associated with mGluR3, then mGluR3 would effectively be an inhibitor of those genes.

Acetylcholine esterase (AChE) is necessary for controlled ACh signaling by catalyzing synaptic breakdown of acetylcholine. Without it, the ACh signal could not degrade and the ACh signaling pathway would be chronically over stimulated and ineffective. Could controlled ACh involving AChE be a general upregulator of the genes associated with AChE in our analysis?

5HT1b is a G-protein coupled serotonin receptor, acts on intracellular signaling, and, like mGluR3, inhibits adenylate cyclase.[11] However, while 5HT1b and mGluR3 share many outputs, their directions of regulation are mutually antagonistic, in contradiction with their shared role in adenylate cyclase inhibition. Keep in mind that our data are derived from whole tissue, not from individual cells. Perhaps the cell types expressing 5HT1b may produce completely different responses than those expressing mGluR3?

GAD67 and GAD65 synthesize the neurotransmitter GABA, of which GRa2 is a receptor, so they are right at the bottom of the GABA signaling cascade. We will cover these genes in more detail in Sec. 3.2.

The bias terms for spinal cord (C_i') and hippocampus ($C_i' + T_i'$) average around zero, and are moderately sized: on the order of the input from a single gene. The difference in bias between the two tissues in the model is on the same order of magnitude, indicating that the tissues are fairly closely related.

The kainate terms K_i' are rather small, which is surprising considering the dramatic and almost instantaneous change in gene expression levels caused by kainate injection. However, two genes show a significant negative influence of kainate: IGF2 (-1.45) and nAChRa3 (-0.87). This leads us to hypothesize that the most direct effect of kainate on gene regulation is on IGF2 and nAChRa3, and that the rest of the changes would be due to reaction of the system to the change in IGF2 and nAChRa3 levels.

The linear model is a very good fit to the original data. A more challenging test is to reconstruct the entire trajectory of the system through state space from scratch: Initialize the gene expression levels to those measured at the very first time point, apply the model once for each time step of the total time span covered by the measurements, updating the simulated expression levels as we go. The linear model indeed simulates almost perfectly the trajectory through state space for all three data sets. Fig. 2 shows the original and reconstructed time series for three representative genes. Interpolated time series (not shown) are nearly indistinguishable from the reconstruction. Analysis of the eigenvectors of the linear system also reveals that the final expression levels are close to fixed points of the system (within 3% for the spinal cord and hippocampus "adult" expression levels, within 9% for the final hippocampus injury expression levels): the linear model settles into an attractor in state space corresponding to the adult expression levels of the real organism.

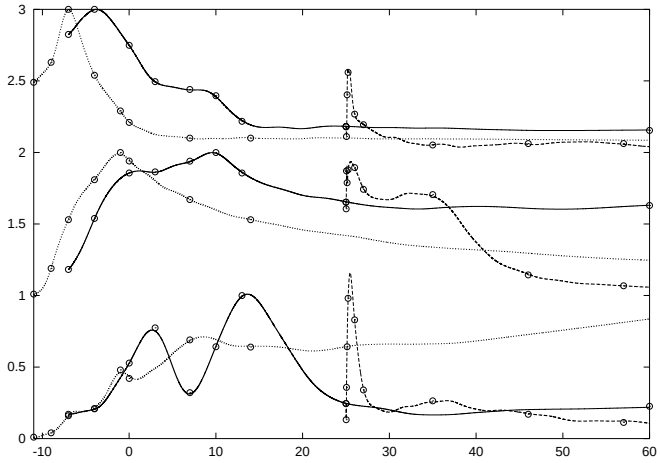

Figure 2: Original (dots) and reconstructed time series (lines) for nestin(top), GRa4 (middle) and aFGF (bottom). Nestin and GRa4 levels are offset by 2.0 and 1.0 respectively. Time is in days from birth (day 0). Dotted line: spinal cord, starting day -11. Solid line: hippocampus development, starting day -7. Dashed line: hippocampus kainate injury, starting day 25.

3.2 Case Study: GAD/GABA Interactions

A graph summarizing the largest weights connecting 43 of the 65 mRNA species is available online,[9] but is unfortunately too large to reproduce here. Such a full scale graph contains a large number of proposed gene interactions that nobody has ever thought of investigating, making it hard to analyze and evaluate. In essence, our knowledge of how the entire system works is too rudimentary to judge whether the overall picture suggested by the linear model makes sense.

Instead, we have chosen to focus on a smaller subsystem: the interaction of GAD (glutamic acid decarboxylase) and GABA-R (γ-amino butyric acid receptors). GABA, synthesized from glutamate by GAD, is a well-known fast-acting synaptic transmitter in the mature CNS. However, it is also thought to play an important role in CNS differentiation during early CNS development.[12] In the rat, two forms of GAD exist, GAD65 and GAD67. There are at least three alternatively spliced transcripts of the precursor mRNA preGAD67: GAD67, G67I86 and G67I8086. GAD expression changes dramatically during development.[13] Earlier models of GAD and GABA-R suggested a positive feedback of GAD and regulation of GABA-R by GAD (via GABA).[14]

The picture presented by the linear model in Fig. 3 is much more detailed. We see indeed a positive autoregulation of GAD65 and preGAD67. There are

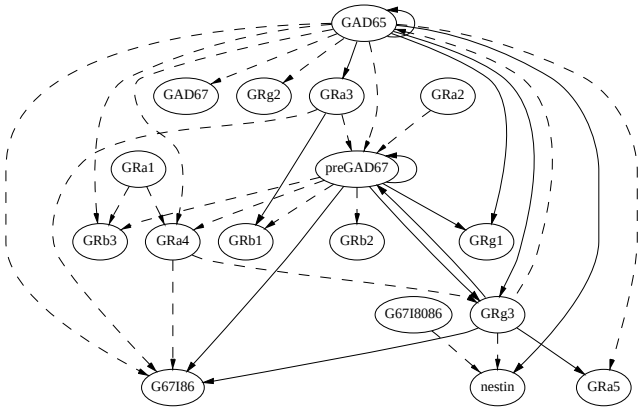

Figure 3: Subgraph with main interactions between GAD and GABA-Receptors. Solid and dashed arrows indicate positive and negative influence respectively.

also some signs of competition between GAD65 and GAD67: GAD65 has a negative effect on all the GAD67 variants, and preGAD67 has (indirectly, via GRg3) a negative effect on GAD65. The strong influence of the pre-mRNA preGAD67, unlikely to have any regulatory function by itself, indicates there could be some regulatory functions shared by the different splicing variants, even though some are not enzymatically active. GAD does indeed seem to affect the GABA receptors, although there are an unexpected number of negative influences. There is more regulation from GAD to GABA-R than vice versa, as predicted.

3.3 Clustering Based On Similar Regulation

We can get an idea of which genes share regulatory inputs by calculating the Euclidean distance between the input vectors (after normalizing their magnitude). This distance measure is directly related to the correlation between the input vectors. Several clusters of genes show very high correlation.

Cluster analysis of the resulting distance matrix using Joe Felsenstein's FITCH program [16] yields the tree in Fig. 4. Several distinct clusters stand out. Whereas G67I8086 and GAD67 are regulated similarly, preGAD67 and G67I86 have different input patterns. This may indicate that a large part of the regulation of GAD67 and its variants occurs post-transcriptionally, i.e. when splicing the pre-mRNA into mRNA. The main differences in regulatory inputs for preGAD67 and GAD67 are 5HT1b (+ for GAD67), GRa2 (+), and

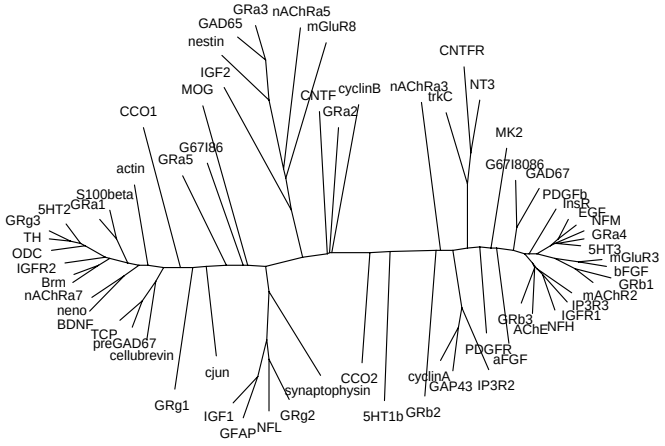

Figure 4: Hierarchical clustering based on input vectors

mGluR3 (−−). GAD65, on the other hand, is regulated very similarly to GRa3 and nestin. Another interesting pair of genes with fairly high correlation is PDGFb/PDGFR, a peptide/receptor pair suggested to be co-regulated.[17]

Not as readily visible in this representation is that there are gene pairs with a very large negative correlation of the input vectors. This may be due to an intermediate factor regulating both genes with different sign, or simply because of a strong negative influence of one gene on the other (see our discussion of causal relationships in Sec. 1). The largest negative correlation is found between Brm and NT3, CNTFR and preGAD67, GRb1 and ODC.

We can also cluster the genes based on their output vectors, indicating similar regulatory functions. Some of the genes with the highest correlation of output vectors are AChE and IGFR2, InsR and NT3, NFM and nAChRa5, mAChR2 and mGluR3. The highest negative correlation is observed between ODC and both NT3 and InsR.

Comparing the output vectors of the six major regulators from Sec. 3.1, several show large positive or negative correlations, indicating a large number of shared outputs. The output vectors of 5HT1b and GRa2, and of mGluR3 and preGAD67, are positively correlated (+0.92 and +0.89). Furthermore, the first two are negatively correlated with the last two (from -0.91 to -0.96). The interpolated time series for mGluR3 and 5HT1b are very similar (correlation of +0.95), so the model might be erroneously adding a large amount of one gene, only to subtract a large amount of a gene with nearly identical

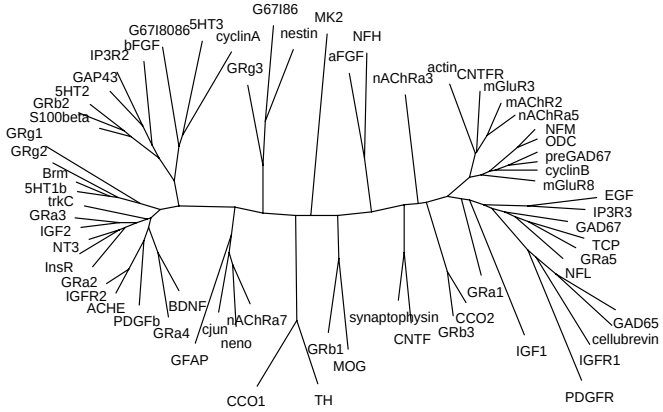

Figure 5: Hierarchical clustering based on output vectors

gene expression levels. The time series for the other genes show much less similarity, so these large correlations between their output vectors could be biologically meaningful, perhaps indicating a mechanism of interlocking up and down regulation.

The two main regulators, AChE and mGluR3, also share a large number of substantial outputs with mutually antagonistic regulation—i.e. all stimulatory outputs of AChE are inhibitory outputs for mGluR3—although the correlation between their output vectors is only -0.67. In addition, five out of the seven neurotransmitter receptors associated with AChE are upregulated (downregulated by mGluR3).

3.4 Functional Categories

When we divide the genes up into functional categories, other interesting patterns emerge. The categories used were: *5HTR* (Serotonin Receptors), *AChR* (Acetylcholine Receptors), *GABA-R* (GABA Receptors), *GluR* (Glutamate Receptors), *ICS* (Intracellular Signaling), *NME* (Neurotransmitter Metabolizing Enzymes, including GAD), *cell cycle*, *glial*, *growth factor*, *insulin & IGF*, *neuronal*, *neurotrophin*, *progenitor*, *synaptic*, *trans-regulation*, and *other*.

NME and *GluR* (mainly mGluR3) are the main input classes, with weights coming from these genes on average more than twice as large as from other genes. Also important are *ICS* (46% larger weights), *5HTR* (45% larger) and *trans regulation* (35% larger). The class of genes with the least influence on other genes in the set is *cell cycle*, followed by *growth factor*, *glial*, *synaptic*,

and *neurotrophin*. We also observed that there is a tendency for genes in one functional class to receive more inputs from genes in the same class.

Notable exceptions to primary regulation by *NME* and *GluR* are *cell cycle*, which receives very little input from *GluR* and a substantial amount from *other* (especially CCO1); and *growth factor*, which gets most input from *ICS* (followed by *5HTR*, *GluR* and *NME* almost equally strong). The input vectors for the *growth factor* genes are also very tightly correlated (PDGFR, MK2, aFGF,PDGFb, EGF, bFGF; all on the right hand side in Fig. 4), indicating that they may share a significant fraction of their regulatory inputs.

4 Discussion

Considering the model presented here has $65 \times 68 = 4420$ parameters, and is generated from a data set of only $65 \times 28 = 1820$ data points before interpolation, there is a danger of overfitting what little data is available. The nonlinear interpolation does impose a significant constraint on the smoothness of the trajectory in between the data points.

One possible method to check whether the resulting model is underconstrained would be to construct a series of similar models by disturbing the input data within the (known) standard deviation for each measurement, and by using different nonlinear interpolation schemes. Comparing these models would tell us how sensitive the results are with respect to small amounts of noise in the input data.

Nevertheless, some features of the model seem to match well with what we assume the underlying system looks like, in terms of sparsity of connections, bias in regulatory function of certain genes, lack of structural genes with significant regulatory function, attractors, known gene interactions, etc. The exact results cited are speculative, but already suggest a number of interesting hypotheses.

The linear modeling approach presented here is very powerful, allowing analysis of a wide range of features of the modeled system, and is able to capture the dynamics of this particular gene regulation system.

The main shortcomings of this approach are 1) the lack of a mechanism to minimize the number of gene interactions, allowing each gene to be modeled by a weighted sum of all other genes, 2) the inherent linearity which can only capture the primary linear components of a presumably nonlinear system, and 3) the need to interpolate non-uniformly spaced data, which gives more weight to widely spaced data points. All three of these problems can be circumvented using a recurrent Neural Network [18] rather than a purely linear model. The equation for such a model is very similar to Eq. 1, except for the addition of

a nonlinear squashing function. The contributions from the regulatory inputs to a gene are still considered additive, but the squashing function allows us to implement a nonlinear dose-response curve, which is a more realistic model for gene regulation (see also the *gene circuit* abstraction of Reinitz and Sharp[7]). A number of well-known training algorithms are available, allowing for desirable features such as reduction of the number of connections, time constants and delays, improving performance with little training data, non-uniformly spaced training data, etc. We are currently working on a model based on this technology, and we expect that the analysis presented here will prove to be an excellent dress rehearsal for a Neural Network based model.

Acknowledgments

This research is funded in part by a grant from the National Science Foundation (grant IRI-9157644). The CNS gene expression time series were generated by X.W, S.F. and R.S. while at the Laboratory of Neurophysiology, NINDS, NIH. We thank the Santa Fe Institute for originating this collaboration.

References

1. X. Wen *et al*, *Proc. Natl. Acad. Sci.* **95**, 334 (1998).
2. Submitted for publication.
3. R. Somogyi *et al*, *Proc. 2nd. World Congress of Nonlin. Anal.* , (1996).
4. P. D'haeseleer *et al*, in *Information Processing in Cells and Tissues*, eds. M. Holcombe and R. Paton (Plenum, New York, 1998).
5. S. Liang *et al*, *Pacific Symposium on Biocomputing* **3**, 18 (1998).
6. A. Arkin *et al*, *Science* **277**, 1275 (1997).
7. J. Reinitz and D. H. Sharp, *Mechanisms of Development* **49**, 133 (1995).
8. J. L. DeRisi *et al*, *Science* **278**, 680 (1997).
9. http://www.cs.unm.edu/~patrik/networks/PSB99/connect.ps
10. L. Prezeau *et al*, *Mol. Pharmacol.* **45**, 570 (1994).
11. P. Schoeffter and D. Hoyer, *Naunyn Schmiedebergs Arch. Pharmacol.* **340**, 285 (1989).
12. J. L. Barker *et al*, *Perspect. Dev. Neurobiol.* **5**, 305 (1998).
13. W. Ma *et al*, *J. Comp. Neurol.* **325**, 257 (1992).
14. R. Somogyi *et al*, *J. Neurosci.* **15**, 2575 (1995).
15. U. Lendahl *et al*, *Cell* **60**, 585 (1990).
16. J. Felsenstein, distributed by the author.
17. http://rsb.info.nih.gov/mol-physiol/BI169/GeNetSomogyi.pdf.
18. B. A. Pearlmutter, *IEEE Trans. Neural Networks* **6**, 1212 (1995).