Abstract of "Analysis of Gene Regulation Networks Using Finite-Field Models"

Humberto Ortiz-Zuazaga

November 7, 2005

1 Summary

Microarrays allow researchers to simultaneously measure the expression of thousands of genes. They give invaluable insight into the transcriptional state of biological systems, and can be important in understanding physiological as well as diseased conditions. However, the analysis of data from many thousands of genes, from only a few replications is very difficult.

The major goal of this proposal is to further develop information theoretic techniques for microarray analysis, and specifically, to develop procedures to cluster gene expression values and determine gene regulatory interactions.

We will use a data set from learning and memory processes in rats to test our procedures.

2 Background

cDNA microarrays are a technique for measuring the abundance of RNA from many thousands of genes simultaneously in an inexpensive experiment (Schena, Shalon, Davis & Brown 1995). They are used extensively for diagnostic purposes, and the data they allow researchers to collect have permitted the study of genome wide interactions among genes. The analysis of microarray data, however, is a difficult task, proving a fruitful area of research in numerous fields. An extensive review is available in (de Jong 2002). This section will attempt to review the literature most relevant to the proposed work.

2.1 Clustering

Clustering of gene expression measurements is an important step in many analysis, most early microarray work performed hierarchal clustering, where genes are successively agglomerated into groups by selecting the two clusters whose average expression values are closest (Eisen, Spellman, Botstein & Brown 1998). It is typical to first cluster genes before trying to determine the gene regulatory network by reverse engineering. Clustering helps reduce the computational resources required to analyze microarray data sets by grouping together many separate genes that demonstrate similar patterns of expression (Akutsu, Miyano & Kuhara 1999). It also can help in determining common functionality or common regulatory elements of genes which cluster together (D'haeseleer, Liang & Somogyi 2000).

2.2 Boolean models and the reverse engineering problem

A series of papers in 1998, 1999 and 2000 defined Boolean network models, reverse engineering, and proved interesting results on the number of experiments required to completely define a Boolean network.

Taking the model definition from (Ideker, Thorsson & Karp 2000), for example, we can describe a genetic network as:

1. A graph consisting of N numbered nodes and, $1 \leq n \leq N$.
2. A set of directed edges between nodes.
3. A Boolean function f_n for each node.

An edge from a node to another represents an influence of the first gene on the expression of the second.

We will additionally define:

An *expression matrix* is a set of measurements (such as those which result from microarray experiments) over the genetic network. From this expression data, the challenge is to reconstruct or reverse engineer the genetic network.

A *gene perturbation experiment* is an expression matrix where some entries correspond to measurements taken when the value of one gene or more are forced to a known state.

The *reverse engineering problem*, then, is to determine the node structure and control functions for a genetic network from an expression matrix or gene perturbation experiment.

Akutsu, Kuahara, Maruyama & Miyano (1998) proved lower and upper bounds on the number of gene perturbation experiments required to completely determine a gene network. The results are discouraging, since in the general case, the problem is shown to be NP-complete. However, in (Liang, Fuhrman & Somogyi 1998), an efficient algorithm for determining the gene network from a set of input-output pairs is developed, assuming that each gene has an indegree in the directed graph that is at most three. This restriction corresponds to saying that at most three genes have an influence on the expression of the target gene. Further research proceeds on the assumption that this indegree is bounded by a small constant. In Akutsu et al. (1999) it is shown that a gene network will be recovered with high probability in only $O(\log n)$ experiments if the indegree is at most two. Ideker et al. (2000) provide an iterative procedure for selecting genes to perturb while determining a genetic network such that the uncertainty in the specification of the model is reduced. After this series of papers, work on these Boolean models was mostly discontinued, biologists objected to the simplicity of the Boolean representation of genes.

It is also important to note that all of these Boolean network papers leave unspecified the manner in which gene expression measurements are converted to Boolean values. For example, Ideker et al. (2000) simply says that gene values will be approximated as high or low and represented by the values 1 or 0.

2.3 Probabilistic models

Probabilistic Boolean networks, PBN, were developed in Shmulevich, Dougherty, Kim & Zhang (2002) to overcome problems encountered in the study of gene expression data with Boolean networks. The principle problem PBNs address is the inherent determinism of Boolean network models. PBN incorporate a stochastic process, to allow for uncertainty in the data, and in the produced

models. PBN can incorporate many Boolean functions for a single gene, selecting among the multiple functions according to a probability that corresponds to how well the function correlates to the data.

2.4 Partial enumeration

In both the Boolean network models and PBN, reverse engineering via partial enumeration of functions as described in (Shmulevich et al. 2002, Liang et al. 1998, Akutsu et al. 1998) requires limiting the number of inputs to each genetic function, usually assuming that between 2 to 4 genes affect the expression of a given gene. This requirement for computational tractability directly conflicts with the evidence that transcriptional networks for higher organisms are significantly more complex (Lemon & Tjian 2000, Merika & Thanos 2001), with even yeast having up to 10 or more transcription factors influencing the expression of a single gene (Lee, Rinaldi, Robert, Odom, Bar-Joseph, Gerber, Hannett, Harbison, Thompson, Simon, Zeitlinger, Jennings, Murray, Gordon, Ren, Wyrick, Tagne, Volkert, Fraenkel, Gifford, & Young 2002).

2.5 Finite field genetic network models

Boolean networks and PBN then share 2 limitations: they can only represent genes as "on" or "off", and they limit the nature of the gene interaction network to ensure computational tractability. Both these problems have been addressed by the formulation of polynomial models over finite fields (Laubenbacher & Stigler 2003). These models allow for a richer variation of gene expression levels, and remove the restrictions on the degree of the genes. These polynomial models, however, are more akin to Boolean network models than to PBN, as they are deterministic, and cannot represent uncertainty in the data or network models.

Several alternative representations and techniques for polynomial models over finite fields have been developed (Aviñó, Green & Moreno 2004, Green 2004, Moreno, Bollman & Aviñó 2002), and Bollman & Orozco (2005) demonstrates that these polynomial models are equivalent to those described in (Laubenbacher & Stigler 2003). This research lead to a series of techniques for error-correction, clustering, and reverse engineering based on finite fields. The current proposal seeks to extend these models, and produce new biological insight from microarray data.

2.6 Microarray experiments

Microarray experiments were performed in the laboratory of Dr. Sandra Peña de Ortiz. Her lab has kindly provided us with data sets for collaborative analysis. The methods described in this proposal were developed for the purpose of analyzing these data sets, but are sufficiently general to analyze any equivalent data set.

The studies described here focus on one cognitive task, conditioned taste aversion (CTA), as a model system for gene expression profiling. CTA, is an associative aversive conditioning paradigm in which pairing gastrointestinal malaise (induced by lithium chloride, LiCl, the unconditioned stimulus) with prior exposure to a novel taste (the conditioned stimulus) may create a strong and long lasting aversion to the novel taste.

CTA lends itself as an excellent model system to study the dynamics of gene regulation in learning and memory because it is a single trial associative learning paradigm, which involves discrete

regions in the brain, including selected amygdala nuclei (Yamamoto, Shimura, Sako, Yasoshima & Sakai 1994, Yasoshima, Shimura & Yamamoto 1995).

Behavioral training of rats in the CTA task prior to collection of the microarray data used for our experiments was done as described in (Ge, Chiesa & Peña de Ortiz 2003).

The gene profiling experiment was replicated five times. Four animals were used per condition for each replicate. Thus, a total of sixteen rats were used per condition. Animals were sacrificed by decapitation at 1, 3, 6, and 24 hours after conditioning. Hybridization, image capture and analysis was similar to the procedures described in (Robles, Vivas, Ortiz-Zuazaga, Felix & Peña de Ortiz 2003). The data set thus obtained (CTA data set) is described in (Chiesa, Ortiz-Zuazaga, Ge & Peña de Ortiz 2000). In summary, the data has two controls, the pre-treatment group and the one hour saline group, and four time points, 1, 3, 6, and 24 hours after conditioning. Each array has 1185 genes, and we have 5 replicates of the arrays.

3 Objectives

As stated in Section 1, our principal goal is to develop new techniques for analyzing microarray data utilizing tools from information theory. These tools have been shown to be applicable to the analysis of microarray expression data. To accomplish our goal we propose the following objectives:

1. Previous Boolean network models assume the values for each gene have been discretized, usually by thresholding, and no errors are present in the discretization. We have seen that using multiple repetitions of an experiment, we can discretize a gene into several values, and use majority logic decoding and other techniques to correct for errors in the microarray image analysis and discretization procedures. Thus we propose **to develop new algorithms and heuristics for clustering and error correction, building on finite field models of gene expression networks, and majority logic decoding.**

2. We have seen that the computational tractability of Boolean and probabilistic Boolean approaches to the reverse engineering problem depend on the assumption that each gene is influenced by a small number of other genes. This assumption is flawed, except perhaps in the simplest of organisms. Multivariate finite field models of gene networks overcome this restriction. We have seen that univariate finite field models are equivalent to multivariate models, and may be simpler to manipulate. This thesis seeks to **develop new algorithms and heuristics for reverse engineering, extending univariate polynomial finite field models to probabilistic models.**

4 Expected outcomes and preliminary results

4.1 Error correction and clustering

We have performed the analysis described above on the CTA data set described in Section 2.6. In this data set, there are 127 consistent genes, which we divide into clusters by grouping together the genes that have the same set of calls in the 1 - 24 hour timepoints. This results in 23 clusters. We focus on the cluster labeled "000+". The consensus of the calls for these genes represents no change over the 1, 3, and 6 hour time points, followed by upregulation at the 24 hour timepoint. This cluster consists of genes whose expression most closely matches the expression profile of CREB.

CREB is a transcription factor which we know to be required for long-term memory (Lamprecht, Hazvi & Dudai 1997).

Two genes in particular caught our interest: Pmch and Calca. Both genes have CRE elements in their upstream regions, meaning they are possible targets of CREB1 regulatory function. Thus these genes exhibit a pattern of expression consistent with the expression of Creb1, have CRE elements upstream of their transcription start site, and seem to have a role in strengthening or creating new synapses. Thus they are strongly implicated as important genes for the formation of memories. Our collaborator, Dr. Sandra Peña de Ortiz, and her students are actively seeking confirmation of these genes' role in CTA. In collaboration with Dr. Moreno, we will confirm the changes in expression of these genes and investigate their role in memory.

4.2 Probabilistic finite field network models

A Probabilistic Finite Field Network (PFFN) is an extension of Probabilistic Boolean Networks (PBN) (Shmulevich et al. 2002) to work over values in finite fields, similar to how finite dynamical systems, as defined in (Laubenbacher & Pareigis 2001) generalize Boolean dynamical systems. In the full proposal, I build a small PFFN, and show how finite fields can be used to encode two particular kinds of models of interest to biologists, a ternary probabilistic network where genes can be unchanged, upregulated or downregulated, and a quaternary probabilistic network where each gene is represented by two Boolean or binary values. In general, we can take a PFFN over $\mathrm{GF}(p^i)$ and split each node into i separate nodes. In the same manner, each predictor may be split into i component parts by taking a basis.

When building transcriptional networks, we may wish to place restrictions on the interactions between genes. For example, we will allow a transcription factor t to act on a gene g only if g has a transcriptional site for t. These types of restrictions can be imposed by restricting the form of allowed predictor functions. Since the available information on transcriptional regulation is incomplete, it is a challenge to incorporate information on allowed, prohibited, and mandatory regulatory interactions, and to do so in an efficient manner.

We will develop tools to perform reverse engineering of PFFN using the model outlined above, and test those tools on the CTA data set.

References

Akutsu, T., Kuahara, S., Maruyama, O. & Miyano, S. (1998), 'Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions', *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms* .

Akutsu, T., Miyano, S. & Kuhara, S. (1999), 'Identification of genetic networks from a small number of gene expression patterns under the Boolean network model', *Pacific Symposium on Biocomputing* **4**, 17–28.

Aviñó, M., Green, E. & Moreno, O. (2004), 'Applications of finite fields to dynamical systems and reverse engineering problems', *Proceedings of the 19th ACM Symposium on Applied Computing - SAC* .

Bollman, D. & Orozco, E. (2005), Finite field models for genetic networks. Preprint.

Chiesa, R., Ortiz-Zuazaga, H. G., Ge, H. & Peña de Ortiz, S. (2000), Gene expression profiling in emotional learning with cDNA microarrays, *in* '40th meeting of the American Society for Cell Biology', San Francisco, California.

de Jong, H. (2002), 'Modeling and simulation of genetic regulatory systems: A literature review', *Journal of Computational Biology* **9**(1), 67–103.

D'haeseleer, P., Liang, S. & Somogyi, R. (2000), 'Genetic network inference: from co-expression clustering to reverse engineering.', *Bioinformatics* **16**(8), 707–726.

Eisen, M., Spellman, P., Botstein, D. & Brown, P. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proceedings of National Academy of Science* **95**, 14863–14867.

Ge, H., Chiesa, R. & Peña de Ortiz, S. (2003), 'Hzf-3 expression in the amygdala after establishment of conditioned taste aversion', *Neuroscience* **120**(1), 1–4.

Green, E. L. (2004), On polynomial solutions to reverse engineering problems. Pre-print.

Ideker, T. E., Thorsson, V. & Karp, R. M. (2000), 'Discovery of regulatory interactions through perturbation: Inference and experimental design', *Pacific Symposium on Biocomputing* **5**, 302–313.

Lamprecht, R., Hazvi, S. & Dudai, Y. (1997), 'cAMP response element-binding protein in the amygdala is required for long- but not short-term conditioned taste aversion memory', *J. Neurosci.* **17**, "8443–8450".

Laubenbacher, R. & Pareigis, B. (2001), 'Equivalence relations on finite dynamical systems', *Advances in Applied Mathematics* **26**, 237–251.

Laubenbacher, R. & Stigler, B. (2003), A computational algebra approach to the reverse engineering of gene regulatory networks. http://arxiv.org/pdf/q-bio.QM/0312026.

Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D., & Young, R. (2002), 'Transcriptional regulatory networks in saccharomyces cerevisiae', *Science* **298**, 799–804.

Lemon, B. & Tjian, R. (2000), 'Orchestrated response: a symphony of transcription factors for gene control', *Genes and Development* **14**(20), 2551–2569.

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.

Merika, M. & Thanos, D. (2001), 'Enhanceosomes', *Curr Opin Genet Dev* **11**(2), 205–208.

Moreno, O., Bollman, D. & Aviñó, M. (2002), 'Finite dynamical systems, linear automata and finite fields', *2002 WSEAS Int. Conf. on System Science Alied Mathematics & Computer Science and Power Engineering Systems* pp. 1481–1483. Also to appear in the International Journal of Computer Research.

Robles, Y., Vivas, P. E., Ortiz-Zuazaga, H. G., Felix, Y. & Peña de Ortiz, S. (2003), 'Hippocampal gene expression profiling in spatial learning', *Neurobiology of Learning and Memory* **80**(1), 80–95.

Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995), 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science* **270**(5235), 467–470.

Shmulevich, I., Dougherty, E. R., Kim, S. & Zhang, W. (2002), 'Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks', *Bioinformatics* **18**(2), 261–274.

Yamamoto, T., Shimura, T., Sako, N., Yasoshima, Y. & Sakai, N. (1994), 'Neural substrates for conditioned taste aversion in the rat', *Behav. Brain Res.* **65**, 1231–137.

Yasoshima, Y., Shimura, T. & Yamamoto, T. (1995), 'Single unit responses of the amygdala after conditioned taste aversion in conscious rats', *Neuroreport* **6**, 2424–2428.