Finite Fields and Microarrays

Humberto Ortiz Zuazaga

humberto@hpcf.upr.edu

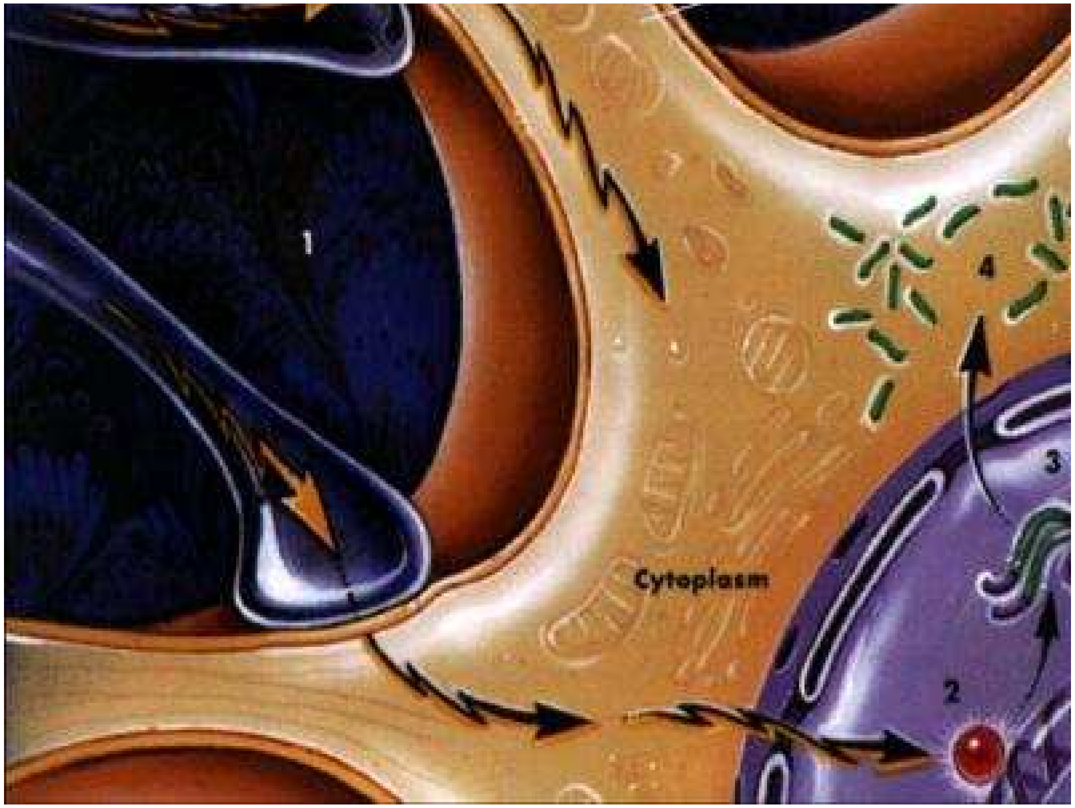
http://www.hpcf.upr.edu/~humberto/

The work I will present is truly a joint, multidisciplinary effort. I have a long, fruitful collaboration with Sandra Peña de Ortiz, the biologist that first got me interested in gene expression (and my wife). María Alicia Aviño Diaz convinced Oscar Moreno that microarrays were a worthwhile item of study for a computer scientist. Dorothy Bollman, Reinhard Laubenbacher, and Carlos Corrada have also contributed results and valuable feedback.

Outline

- Post-genome biology
- Microarrays
- Genetic networks
- Reverse engineering
- Boolean genetic network models
- Finite fields are better Booleans
- Summary and conclusions

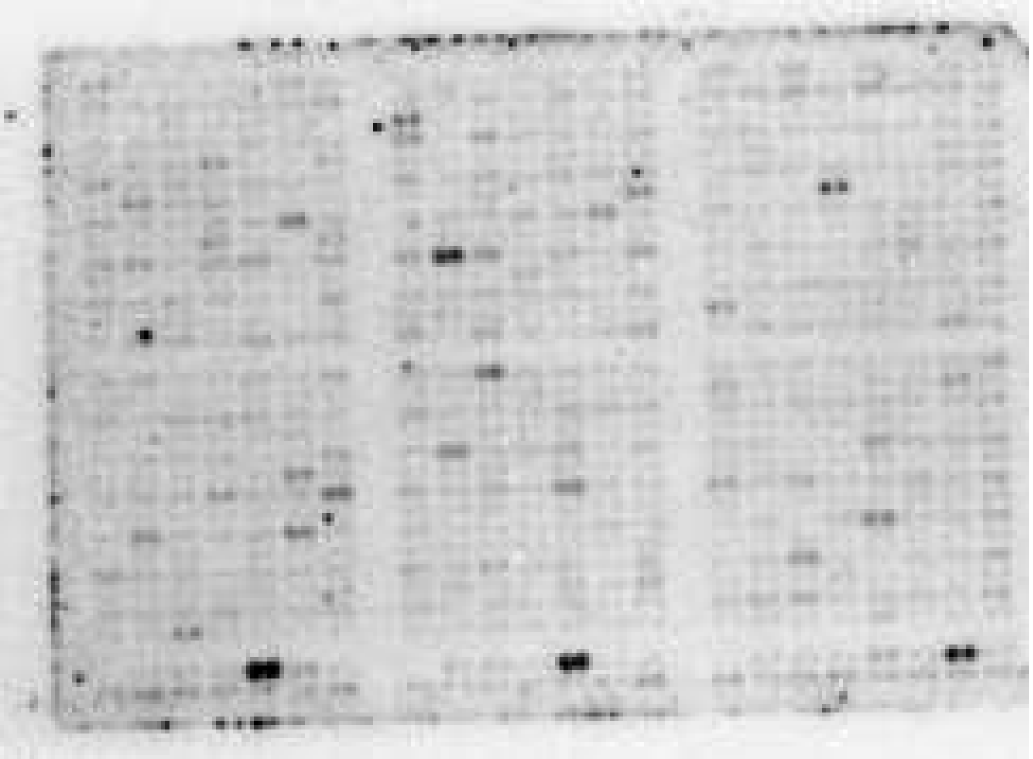
A Model Cell



Modern Molecular Biology in 2 minutes: 1, signals are received at the cell surface, and travel eventually to the nucleus, 2 where transcription factors cause the signal to be converted into a change in expression of a gene. 3 The gene products are converted to proteins in the cytoplasm, 4 where they can now effect further changes in the cell.

Post Genome Biology

or, “I’ve got all the genes, now what do I do with them?”



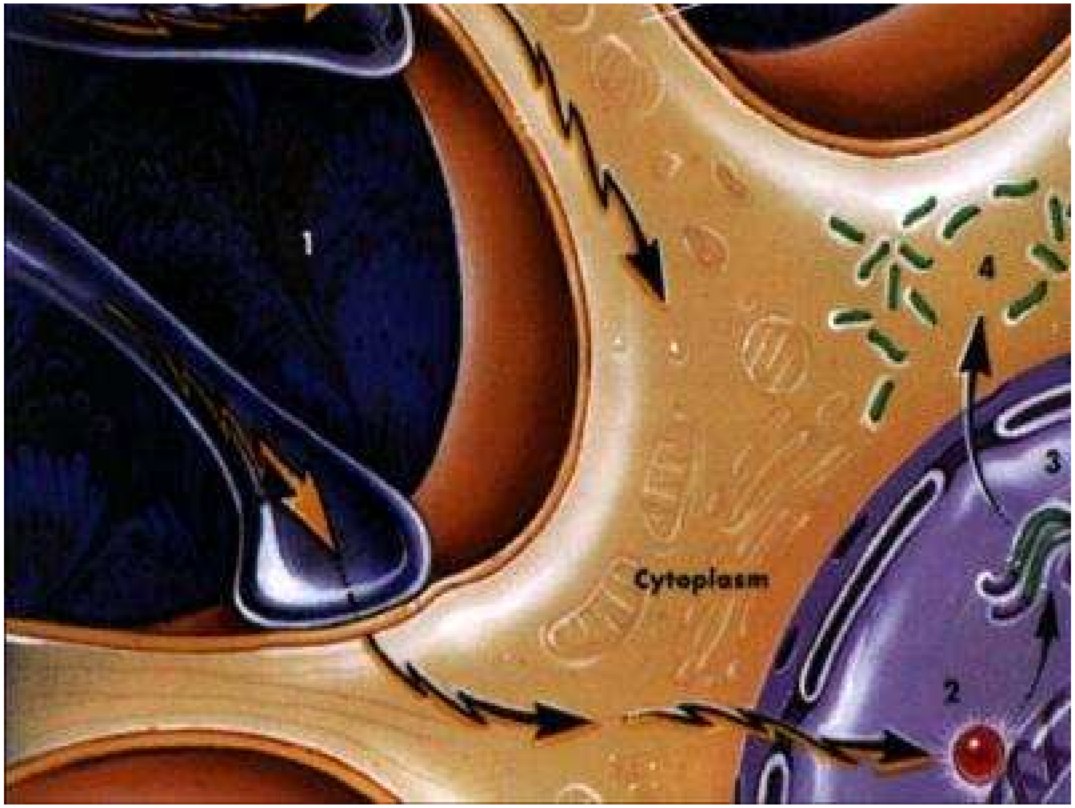
A sample microarray image from Sandra Peña's lab, of 588 genes, spotted in pairs on a nylon membrane, labeled with a radioactive probe, and imaged with a phosphoimager. This data are from an experiment measuring the effect of suppressing CREB, a transcription factor which is required for the formation of memories.

Kida S, Josselyn SA, de Ortiz SP, Kogan JH, Chevere I, Masushige S, Silva AJ. CREB required for the stability of new and reactivated fear memories. Nature Neuroscience. 2002 Apr;5(4):348–55.

One single experiment can measure the levels of expression of all these genes simultaneously.

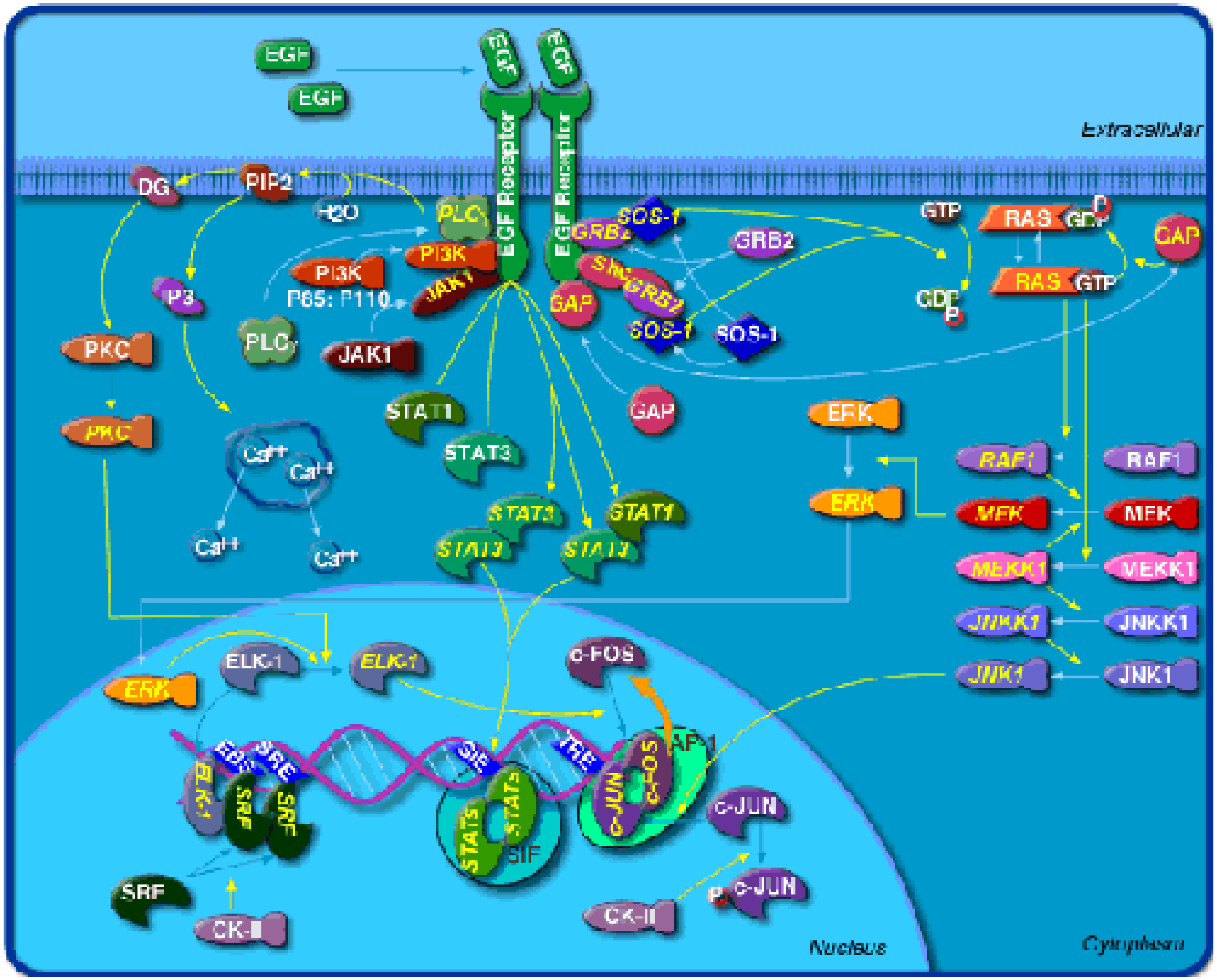
Once an organism is sequenced, there is no technical barrier to developing microarrays with every possible gene from that organism (already done for approx $6,000$ genes in yeast).

The Cell Again



Returning to our sample cell, a microarray experiment only measures or alters the output of step 2 in this diagram, the expression of genes. We wish to infer step 1 through 4.

Genetic Networks



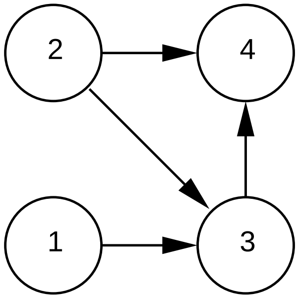
Things start to get complicated, or my “scare the mathematician” slide. This is a portion of a real genetic network regulating the expression of one gene. Again, a microarray experiment only measures the expression of the gene, and we must try to infer the interactions among all these components.

Reverse Engineering Genetic Networks

- Input:
  - A set of genes
  - A set of gene expression measurements
- Output:
  - A set of *control functions* by which some genes control others

Reverse engineering is done by performing a set of experiments, where genes or stimuli are manipulated to be in a known state. By observing how the rest of the gene expression is altered, we can begin to determine how the expression of each gene depends on the others.

Boolean Genetic Networks



$$
\begin{aligned}
f_1 &= 1 \\
f_2 &= 1 \\
f_3 &= x_1 \wedge x_2 \\
f_4 &= x_2 \wedge \neg x_3
\end{aligned}
$$

A small sample genetic network, from Ideker et al.. Nodes are genes, edges are regulatory interactions. The Boolean control functions on the right completely describe the behavior of the network. Finding these functions is the goal of reverse engineering.

Boolean Genetic Network Model

In [3] we define Boolean genetic network model (BGNM):

- A *Boolean variable* takes the values 0, 1.
- A *Boolean function* is a function of Boolean variables, using the operations $\wedge$, $\vee$, $\neg$.

A *Boolean genetic network model* (BGNM) is:

- An n-tuple of Boolean variables $(x_1, \ldots, x_n)$ associated with the genes
- An n-tuple of Boolean control functions $(f_1, \ldots, f_n)$, describing how the genes are regulated

Reverse Engineering Boolean Networks

- Akutsu, S. Kuahara, T. 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 (SODA 98), H. Karloff, ed. ACM Press.

- Ideker, T.E., Thorsson, V., and Karp, R.M. 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. Pacific Symposium on Biocomputing 5:302-313.

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

There has been extensive work on Boolean genetic networks, since (at least) 1996. The problem of determining if a given assignment to all the variables is consistent with a given gene network was shown to be NP-complete by Akutsu et al. (by reduction from 3-SAT). They also provide bounds on the number of experiments required to reverse engineer a gene network under several assumptions.

In the worst case, $2^{(n-1)/2}$ experiments are needed, but if the indegree of each node (the genes that affect our target gene) is bound by a constant D, the cost is $O(n^{2D})$. Under this assumption, Ideker et al and Liang et al provide effective procedures for reverse engineering, assuming any gene may be set to any value.

Boolean Bugs

- Boolean variables can only represent all-or-none effects
- Boolean algebra cannot be generalized to multiple states

Finite fields represent an alternative algebraic structure to substitute Booleans. Our research seeks to characterize genetic networks based on these fields.

Biologists dislike Boolean networks because of the first limitation.

The second limitation irritates mathematicians.

Because of these limitations, we sought to generalize BGNM to models with more states, and other types of interactions besides Boolean logic. Finite fields, a subject of considerable study in recent years because of important applications in communications and coding theory, were our choice for extending genetic network models.

Finite Fields

A *finite field* $\{F, +, \cdot\}$ is a finite set F, and two operations $+$ and $\cdot$ that satisfy the following properties:

- $\forall a, b \in F$, $a + b \in F$, $a \cdot b \in F$
- $\forall a, b \in F$, $a + b = b + a$, $a \cdot b = b \cdot a$
- $\forall a, b, c \in F$, $a + (b + c) = (a + b) + c$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
- $\forall a, b, c \in F$, $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$
- $\exists 0, 1 \in F$, $a + 0 = 0 + a = a$, $a \cdot 1 = 1 \cdot a = a$
- $\forall a \in F$, $\exists (-a) \in F$ s.t. $a + (-a) = (-a) + a = 0$
  $\forall a \neq 0 \in F, \exists a^{-1} \in F$ s.t. $a \cdot a^{-1} = a^{-1} \cdot a = 1$

The field is closed under both operations, both operations are commutative and associative, and the distributive law holds. There are additive and multiplicative identities and inverses.

The real and rational numbers are fields with an infinite number of elements. A finite field has the same properties as the rational numbers, over a finite set. In particular, we can add, subtract, multiply and divide any element by any other.

The World's Smallest Finite Field

The integers 0 and 1, with integer addition and multiplication modulo 2 form the finite field $Z_2 = \{\{0,1\},+,\cdot\}$.

The operators $+$ and $\cdot$ are defined as follows:

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

From the function tables, you can directly verify that the 6 properties of finite fields hold.

Products of Sums and Sums of Products

We can realize any Boolean function as an expression over Z_2:

$$\begin{aligned} X \wedge Y &= X \cdot Y \\ X \vee Y &= X + Y + X \cdot Y \\ \neg X &= 1 + X \end{aligned}$$

See for example Patterson and Hennessy's Computer Organization and Design for the realization of Boolean functions using products and sums.

Note also that $+$ corresponds to the exclusive or (**xor**) Boolean function, so all Booleans functions can be realized with **and** and **xor**.

Finite Field Genetic Networks

Any BGNM can be converted into an equivalent model over Z_2 by realizing the boolean functions as sums-of-products and products-of-sums. We now have a *finite field genetic network* (FFGN):

- An n-tuple of variables over Z_2, $(x_1, \ldots, x_n)$ associated with the genes

- An n-tuple of functions over Z_2, $(f_1, \ldots, f_n)$, describing how the genes are regulated

This result was submitted to the Fifth Annual Conference on Computational Molecular Biology (RECOMB 2003), where I will present a poster on FFGN (the poster was accepted last week!).

The BGNM and FFGN over Z_2 are exactly equivalent, so the complexity and reverse engineering results from the previous papers still apply.

This is the FFGN over Z_2, but these are generalizable to any finite field. We will see what kinds of finite fields we can construct.

A Field With 3 Elements

We can construct a field $\{\{-1, 0, 1\}, +, \cdot\}$, with $+$, $\cdot$ defined as:

$+$	-1	0	1
-1	1	-1	0
0	-1	0	1
1	0	1	-1

$\cdot$	-1	0	1
-1	1	0	-1
0	0	0	0
1	-1	0	1

If you consider a microarray experiment that measures changes from a reference condition, you can have 3 states: upregulated (1), no change (0), or downregulated (-1) these are biologically meaningful classifications which a biologist would not argue with.

A peculiar property of this field is that upregulated + upregulated = downregulated ($1 + 1 = -1$). This is unintuitive for biologists (and me).

Finite Fields are Better Booleans

For any prime integer p, there is a finite field (Z_p) consisting of the integers modulo p with the operations $+$, $\cdot$ modulo p as described above.

As well, we can define a FFGN over Z_p:

- An n-tuple of variables over Z_p, $(x_1, \ldots, x_n)$ associated with the genes
- An n-tuple of functions over Z_p, $(f_1, \ldots, f_n)$, describing how the genes are regulated

Compared to BGNM, FFGN allow us to model genes as variables with many states, and allow instead of simple Boolean logic, the modeling of *additive* $(+)$ and *synergistic* $(\cdot)$ control functions.

We do not know the complexity of reverse engineering FFGN with more than 2 states, but the procedures described in Ideker et al and Liang et al are still valid.

Other Current Results

In papers we have submitted, we show several other results following from the idea of using finite fields:

- BGNM are equivalent to Finite Dynamical Systems
- New procedures for reverse engineering from time-series data
- BGNM are also Finite State Systems
- A framework for correcting errors in microarray data based on deviations from a known genetic network.

See the references section at the end. Each of these results builds upon the basic observation that BGNM can be recast as models over a finite field. Each extends our results in a slightly different direction, and brings in different analytical methods.

Future Work

- Find the complexity of the reverse engineering problem for FFGN
- Test our procedures on real and simulated microarray expression data
- Develop models that capture more biological knowledge

Summary

- BGNM have limitations
- FFGN over Z_2 are equivalent to BGNM
- FFGN over Z_p allow many more states, and different types of control functions
- FFGN over Z_p also allow a large body of modern knowledge on algebra, communications theory, electrical engineering and computer science to be employed to analyze microarray experiments

Publications

1. Ortiz-Zuazaga, H., Aviño-Diaz, M. A., Laubenbacher, R., Moreno O. Finite fields are better Booleans. Refereed abstract, poster to be presented at the Seventh Annual Conference on Computational Molecular Biology (RECOMB 2003), April 10–13, 2003, Germany.

2. Moreno, O., Ortiz-Zuazaga, H., Corrada Bravo, C. J., Aviño-Diaz, M. A., Bollman, D. A finite field deterministic genetic network model. Submitted to the fifteenth conference on Applied Algebra and Error Correcting Codes (AAECC-15), May 12–16, 2003, Toulouse, France.

3. Ortiz-Zuazaga, H., Aviño-Diaz, M. A., Corrada Bravo, C. J., Laubenbacher, R., Peña-de-Ortiz, S., Moreno O. Applications of finite fields to the study of microarray expression data. Submitted to the 11th International Conference on Intelligent Systems for Molecular Biology (ISMB 2003), June 29 to July 3, 2003, Brisbane, Australia.