Bioinformatics and Computational Biology

Humberto Ortiz Zuazaga

University of Puerto Rico
High Performance Computing facility

July 16, 2009

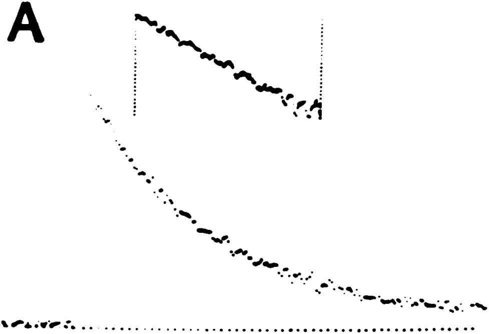
Bioinformatics

"The creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems posed by or inspired from the management and analysis of biological data." — Wikipedia

Computational biology

The application of computers to the collection, analysis, and presentation of biological information.

Electrophysiological data collection

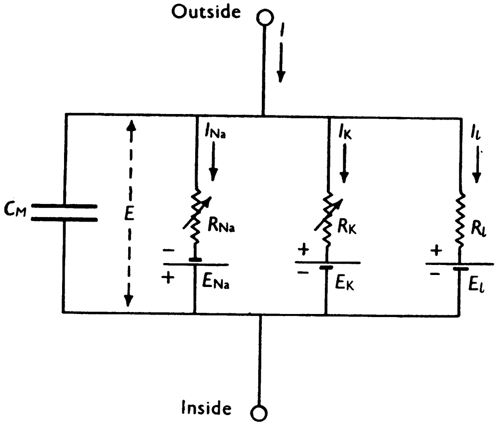


Steinacker A, Zuazaga DC. Changes in neuromuscular junction endplate current time constants produced by sulfhydryl reagents. Proc Natl Acad Sci U S A. 1981 Dec;78(12):7806–7809.

Data collection system

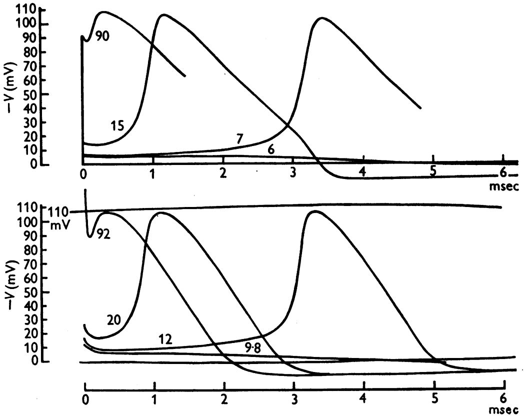
Digital Equipment Corporation (DEC) PDP-11. Replaced high speed camera pictures of oscilloscope followed by manual measurement of trace heights encoded on a deck of punched cards for processing by IBM mainframe in Facundo Bueso.

Electrophysiological simulation



A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol. 1952 August 28; 117(4): 500–544.

Electrophysiological verification



Computed action potentials on top, experimental action potentials on bottom. Awarded the 1963 Nobel Prize in Physiology or Medicine.

Moore’s law

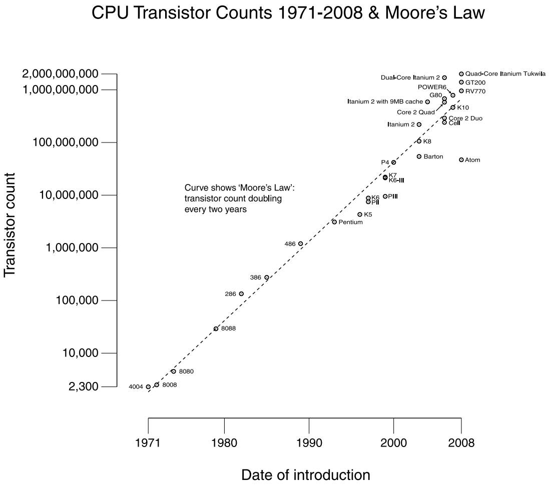
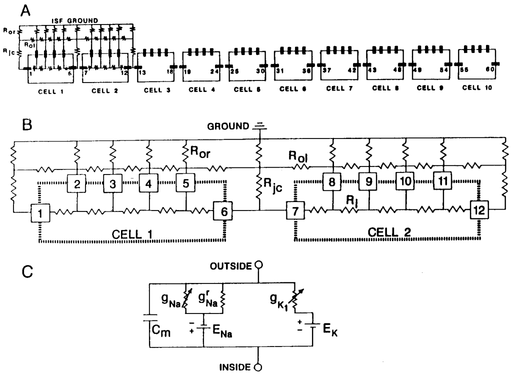


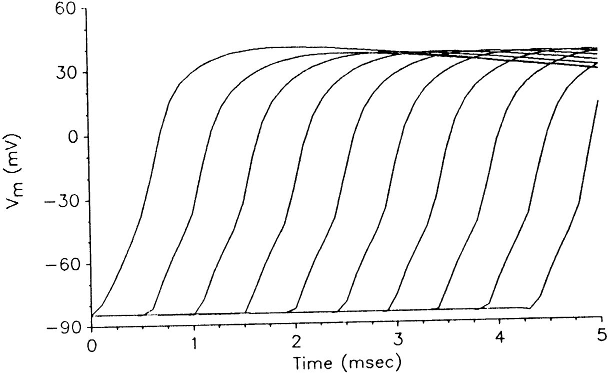
Image from Wikimedia commons by Wgsimon, used with permission.

Larger scale simulations

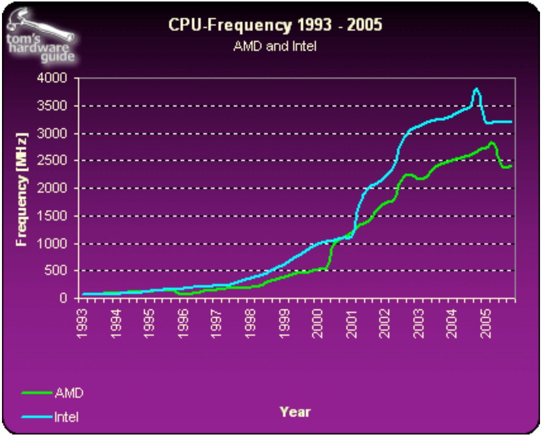


N. Sperelakis, **H. Ortiz-Zuazaga**, and J. B. Picone. Fast conduction in the electric field model for propagation in cardiac muscle. Innov. et Tech. en Biol. et Med., 12(4):404-414, 1991.

Larger scale results

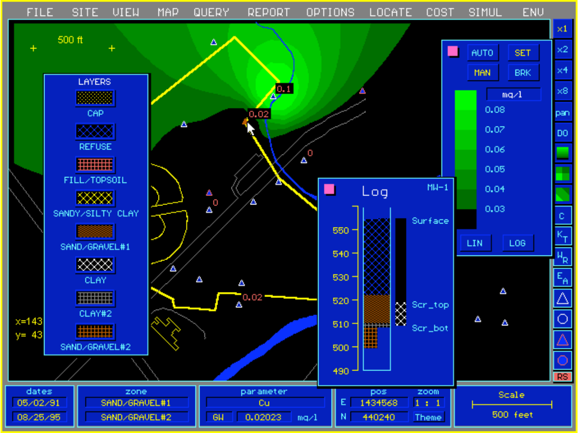


The end of Moore's law



Where's my 4 GHz processor?

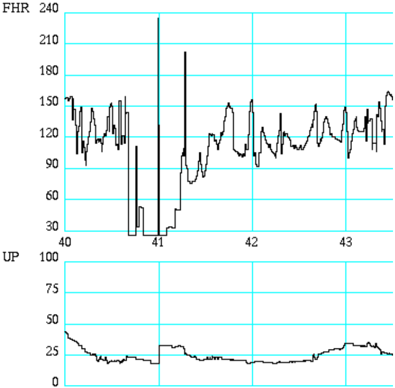
Simulation of groundwater contamination



dates	zone	parameter			pos		zoom	Scale
05/02/91	SAND/GRAVEL#1	Cu			E	1434568	1 : 1	
08/25/95	SAND/GRAVEL#2	GW	0.02023	mg/l	N	440240	Theme	500 feet

A GRACE interface for GRASS. John Franco and **Humberto Ortiz-Zuazaga.** U.S. Army Corps of Engineers, $75,000, 1994–1995.

Neural network processing of cardiotocograms

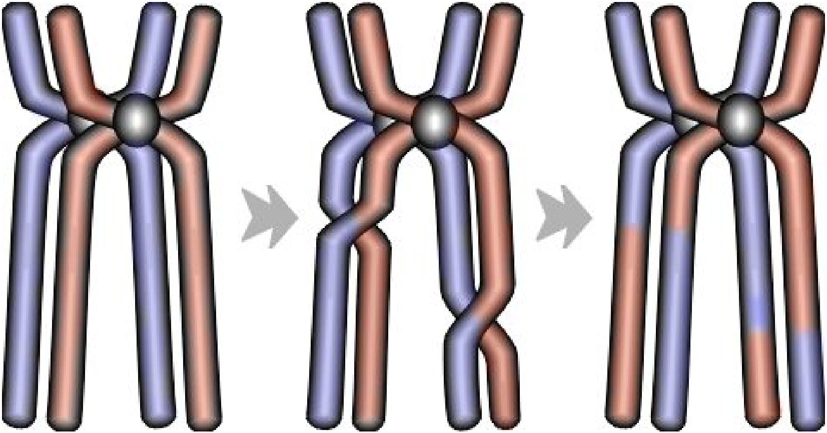


B. E. Rosen, D. Soriano, T. Bylander, and **H. Ortiz-Zuazaga.** “Training Neural Networks to recognize Artifacts and Decelerations in Cardiotocograms.” AAAI Symposium on Artificial Intelligence in Medicine. pp. 149–153, 1996.

Genetic Mapping

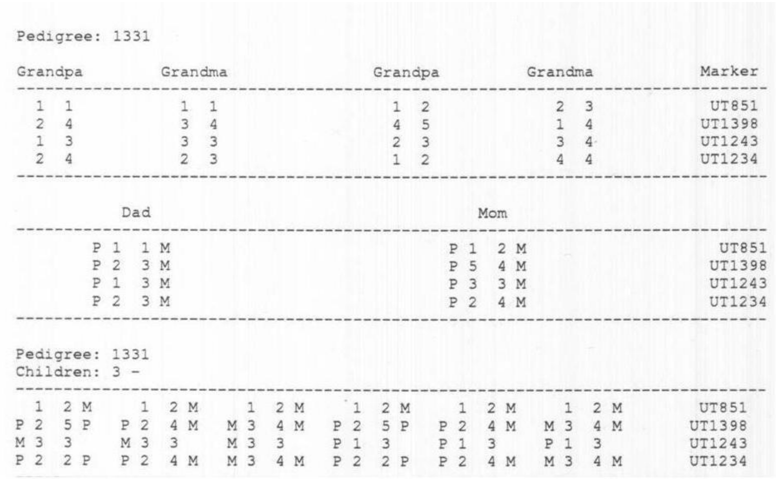
- Goal: The determination of orders and distances among markers on a chromosome based on the observed patterns of inheritance of the alleles of the markers in three generation pedigrees.
- Problem: For a variety of reasons the genotypic information is not complete, and not all crosses in human pedigrees are informative. In addition, the time required to order markers grows exponentially with the number of markers.
- Solution: Only use "good" markers to make maps. Biologists already have a notion of a "framework" map, a map of a subset of the markers which has very high odds against inversion of adjacent markers.

Meiotic breakpoints



From http://www.stanford.edu/group/Urchin/

A genotyped pedigree



```
Pedigree: 1331

Grandpa        Grandma               Grandpa         Grandma           Marker
------------------------------------------------------------------------------
  1  1           1  1                  1  2            2  3            UT851
  2  4           3  4                  4  5            1  4           UT1398
  1  3           3  3                  2  3            3  4           UT1243
  2  4           2  3                  1  2            4  4           UT1234
------------------------------------------------------------------------------

           Dad                                  Mom
------------------------------------------------------------------------------
        P 1  1 M                             P 1  2 M                    UT851
        P 2  3 M                             P 5  4 M                   UT1398
        P 1  3 M                             P 3  3 M                   UT1243
        P 2  3 M                             P 2  4 M                   UT1234
------------------------------------------------------------------------------

Pedigree: 1331
Children: 3 -
------------------------------------------------------------------------------
  1  2 M     1  2 M     1  2 M     1  2 M     1  2 M     1  2 M        UT851
P 2  5 P   P 2  4 M   M 3  4 M   P 2  5 P   P 2  4 M   M 3  4 M       UT1398
M 3  3     M 3  3     M 3  3     P 1  3     P 1  3     P 1  3         UT1243
P 2  2 P   P 2  4 M   M 3  4 M   P 2  2 P   P 2  4 M   M 3  4 M       UT1234
```

Counting obligate breaks as an estimate of genetic distance

A simple estimate of the genetic distance between two markers is the number of observed recombinations between the markers in the data set. For the first two markers in our sample pedigree we would have:

UT851	U	M	U	M	U	M	U	M	U	M	U	M
UT1398	P	P	P	M	M	M	P	P	P	M	M	M
Breaks		1						1				

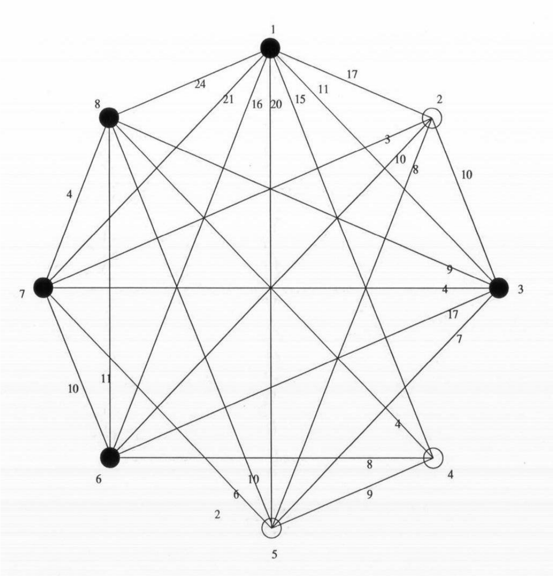
for a total of 2 breaks.

This technique based on counting the number of recombinations is known as meiotic breakpoint analysis (BPA).

Selecting genetic markers with `wclique`

- Each marker becomes a node of a graph.
- The weight of the node is the total count of P and M phases for this marker.
- Two nodes in the graph are connected by an edge whose weight is the number of breaks between the corresponding markers.

A small distance graph



Finding framework markers is a graph problem

- Finding a good set of framework markers is now a graph problem: find a set of nodes with maximal weight where all the nodes are connected by an edge of weight e or higher.
- This graph problem is called Maximal Weighted Clique (MWC).

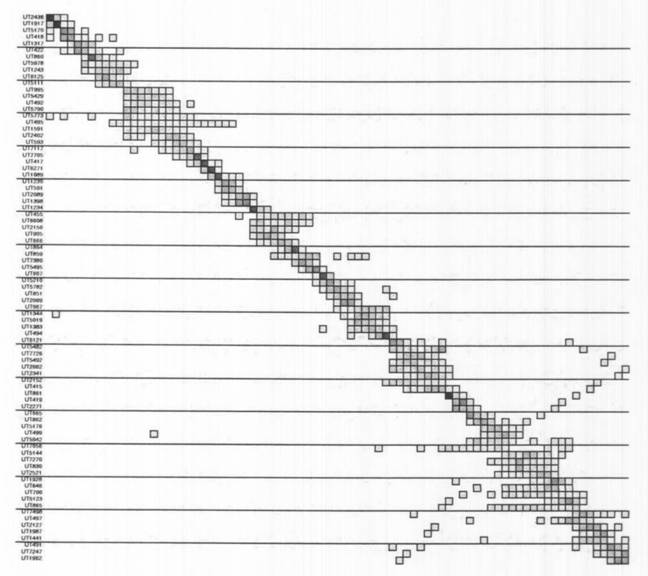
The maximal weighted clique problem is NP-complete

- The MWC is a well known graph problem, extensively studied in computer science. Unfortunately, it belongs to the class of NP-complete problems, for which there is unlikely to be an efficient algorithm.
- Building a linear map by ordering genetic markers so as to minimize the number of recombination events in a set of gametes can also be cast as a graph problem, the traveling salesman problem (TSP), which is also NP-complete.

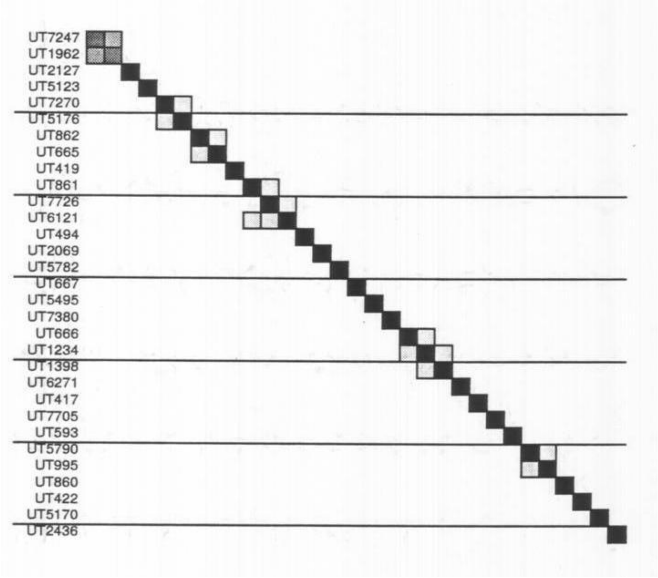
But I still need a map

- Exact algorithms can work on small sets of markers.
- Local search techniques can find near optimal solutions for some of these problems, at the cost of not knowing if an optimal solution was ever found. The best heuristics for TSP can find a solution with 1.05 times the optimal cost.
- A change in the formulation of the problems can enable other algorithms to be used. For example, if the data had no errors, was complete, and no double recombination events occurred, ordering genetic markers would be equivalent to the consecutive ones problem (C1P) for which there are linear time algorithms.

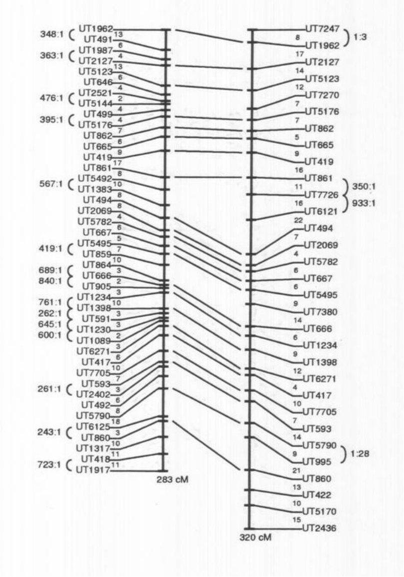
Searls plot of unselected markers



Searls plot of `wclique`-selected markers



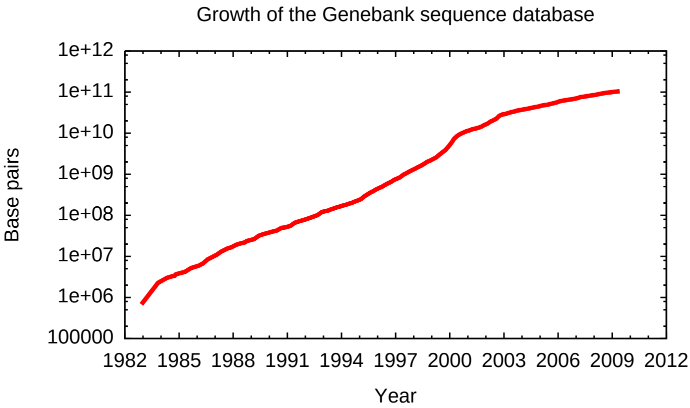
Comparison of MLA maps of hand-selected and `wclique`-selected markers



References

1. **H. Ortiz-Zuazaga,** and R. Plaetke. Screening genetic markers with the maximum weighted clique method. Abstract presented at Genome Mapping and Sequencing. Cold Spring Harbor, May 1997.
2. S.L. Naylor, R. Plaetke, **H. Ortiz-Zuazaga,** P. O'Connell, B. Reus, X. He, R. Linn, S. Wood, and R.J. Leach. Construction of Framework and Radiation Hybrid Maps of Chromosomes 3 and 8. Abstract presented at Genome Mapping and Sequencing. Cold Spring Harbor, NY, May 1997.

“Moore’s law” for sequence data

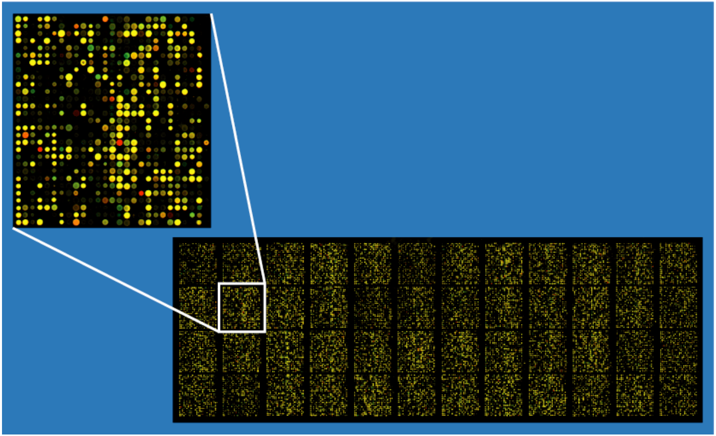


From the June 15 2009 NCBI-GenBank Flat File Release 172.0

Gene expression networks

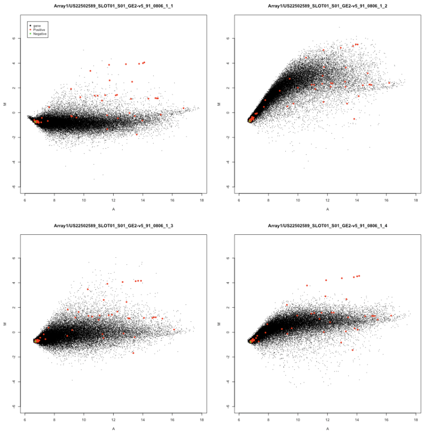
- Complete genomes available for several species.
- 40,000 human genes, many already sequenced.
- microarrays can measure expression levels for ALL GENES in a single assay.

Microarray image



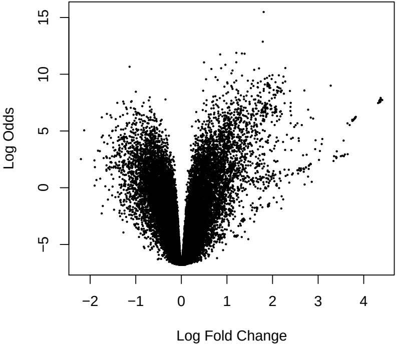
Reproduced from www.molecularstation.com

Microarray data



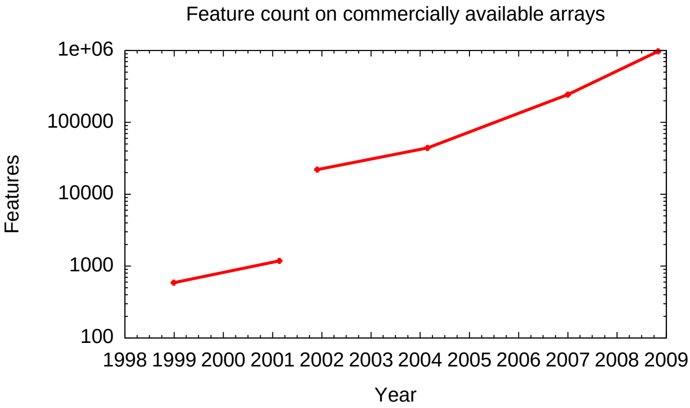
Raw log ratio vs log intensity for two color microarrays.

Microarray analysis



Find the differentially expressed genes.

“Moore’s law” for microarrays



Boolean Genetic Network Model

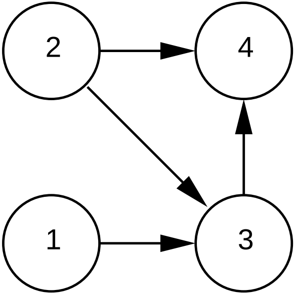
We define Boolean Genetic Network Model (BGNM):

- 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

Boolean genetic networks



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

Previous results on Boolean networks

- Determining if a given assignment to all the variables is consistent with a given gene network was shown to be NP-complete in [1] (by reduction from 3-SAT).
- In the worst case, $2^{(n-1)/2}$ experiments are needed
- If the indegree of each node (the genes that affect our target gene) is bound by a constant D, the cost is $O(n^{2D})$.
- For low D, [2] and [3] provide effective procedures for reverse engineering, assuming any gene may be set to any value.

Reverse engineering Boolean networks

1. 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.
2. 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.
3. 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.

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

Finite field equivalents to the Boolean operators

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

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

Finite field genetic networks

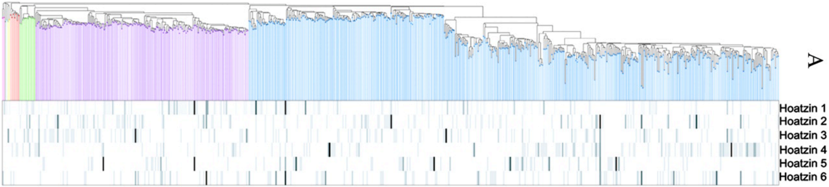
Any BGNM can be converted into an equivalent model over Z_2 by realizing the Boolean functions as sums-of-products and products-of-sums, then converting the Booleans to Z_2. We now have a *finite field genetic network* (FFGN):

- 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

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. **Humberto Ortiz-Zuazaga,** Sandra Peña de Ortiz, Oscar Moreno de Ayala. Error Correction and Clustering Gene Expression Data Using Majority Logic Decoding. Proceedings of The 2007 International Conference on Bioinformatics and Computational Biology (BIOCOMP'07), Las Vegas, Nevada, June 25–28, 2007.
3. **Humberto Ortiz Zuazaga,** Tim Tully, Oscar Moreno. Majority logic decoding for probe-level microarray data. Proceedings of BIOCOMP'08 — The 2008 International Conference on Bioinformatics and Computational Biology, Las Vegas, Nevada, July 13–17, 2008.

Molecular phylogeny



Filipa Godoy-Vitorino, Ruth E. Ley, Zhan Gao, Zhiheng Pei, **Humberto Ortiz-Zuazaga,** Luis R. Pericchi, Maria A. Garcia-Amado, Fabian Michelangeli, Martin J. Blaser, Jeffrey I. Gordon, Maria G. Dominguez-Bello. Bacterial Community in the Crop of the Hoatzin, a Neotropical Folivorous Flying Bird. Applied and Environmental Microbiology, October 2008, p. 5905–5912, Vol. 74, No. 19. doi:10.1128/AEM.00574-08