Polynomial dynamical systems over finite fields, with applications to modeling and simulation of biological networks.

IMA Workshop on
Applications of Algebraic Geometry in
Biology, Dynamics, and Statistics
March 6, 2007

Reinhard Laubenbacher
Virginia Bioinformatics Institute
and Mathematics Department
Virginia Tech
Polynomial dynamical systems

Let $k$ be a finite field and $f_1, \ldots, f_n \in k[x_1, \ldots, x_n]$

$$f = (f_1, \ldots, f_n) : k^n \to k^n$$

is an $n$-dimensional polynomial dynamical system over $k$.

Natural generalization of Boolean networks.

Fact: Every function $k^n \to k$ can be represented by a polynomial, so all finite dynamical systems $k^n \to k^n$ are polynomial dynamical systems.
Example

\[ k = F_3 = \{0, 1, 2\}, \quad n = 3 \]
\[ f_1 = x_1 x_2^2 + x_3, \]
\[ f_2 = x_2 + x_3, \]
\[ f_3 = x_1^2 + x_2^2. \]
Motivation: Gene regulatory networks

“[The] transcriptional control of a gene can be described by a discrete-valued function of several discrete-valued variables.”

“A regulatory network, consisting of many interacting genes and transcription factors, can be described as a collection of interrelated discrete functions and depicted by a wiring diagram similar to the diagram of a digital logic circuit.”

Karp, 2002
The segment polarity network is a robust developmental module

George von Dassow, Eli Meir, Edwin M. Munro & Garrett M. Odell

University of Washington, Department of Zoology, Box 351800, Seattle, Washington 98195-1800, USA

Nature 406 2000
frequent; as this search involved 48 parameters, on average a random choice of parameter value has roughly a 90% chance of being compatible with the desired behaviour (0.9^{48} is \sim 1/200). This holds even though most parameters range over several orders of magnitude. For comparison, if the model tolerated variation in the average parameter over 10% of its 100- or 1,000-fold range (a wildly optimistic expectation for a human-engineered electronic circuit), random search would find only one solution in 10^{48} samples.
The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*

Réka Albert*, Hans G. Othmer

*School of Mathematics, University of Minnesota, 127 Vincent Hall, Minneapolis, MN 55455, USA*

Received 28 June 2002; received in revised form 6 December 2002; accepted 9 December 2002
Motivation (2): a mathematical formalism for agent-based simulation

- Example 1: Game of life
- Example 2: Large-scale simulations of population dynamics and epidemiological networks (e.g., the city of Chicago)

Need a mathematical formalism.
Variables $x_1, \ldots, x_n$ with values in $k$.

$(s_1, t_1), \ldots, (s_r, t_r)$ state transition observations with $s_j, t_j \in k^n$.

**Network inference:**

Identify a collection of “most likely” models/dynamical systems

$$f=(f_1, \ldots, f_n): k^n \to k^n$$

such that $f(s_j)=t_j$. 
Important model information obtained from $f=(f_1, \ldots, f_n)$:

- The “wiring diagram” or “dependency graph”
  
  directed graph with the variables as vertices; there is an edge $i \rightarrow j$ if $x_i$ appears in $f_j$.  

- The dynamics
  
  directed graph with the elements of $k^n$ as vertices; there is an edge $u \rightarrow v$ if $f(u) = v$.  

The Hallmarks of Cancer  Hanahan & Weinberg (2000)
The model space

Let $I$ be the ideal of the points $s_1, \ldots, s_r$, that is,

$$I = \langle f \in k[x_1, \ldots, x_n] \mid f(s_i) = 0 \text{ for all } i \rangle.$$

Let $f = (f_1, \ldots, f_n)$ be one particular feasible model. Then the space $M$ of all feasible models is

$$M = f + I = (f_1 + I, \ldots, f_n + I).$$
Problem: Given data \((s_i, t_i), i=1, \ldots , r\),
(a collection of state transitions for one node in the network),
find all \textit{minimal} (wrt inclusion) sets of variables \(y_1, \ldots , y_m \in \{x_1, \ldots , x_n\}\) such that
\[(f + I) \cap k[y_1, \ldots , y_m] \neq \emptyset.\]
Each such minimal set corresponds to a minimal wiring diagram for the variable under consideration.
The “minimal sets” algorithm

For \( a \in k \), let \( X_a = \{ s_i \mid t_i = a \} \).
Let \( X = \{ X_a \mid a \in k \} \).

Then
\[
 f^0 + I = M = \{ f \in k[x] \mid f(p) = a \text{ for all } p \in X_a \}. 
\]

Want to find \( f \in M \) which involves a minimal number of variables, i.e., there is no \( g \in M \) whose support is properly contained in \( \text{supp}(f) \).
The algorithm

Definitions.

• For $F \subset \{1, \ldots, n\}$, let
  \[ R_F = k[x_i \mid i \notin F]. \]

• Let $\Delta_X = \{F \mid M \cap R_F \neq \emptyset\}$.

• For $p \in X_a$, $q \in X_b$, $a \neq b \in k$, let
  \[ m(p, q) = \prod_{i \neq q_i} x_i. \]

Let $M_X = \text{monomial ideal in } k[x_1, \ldots, x_n]$ generated by all monomials $m(p, q)$ for all $a, b \in k$.

(Note that $\Delta_X$ is a simplicial complex, and $M_X$ is the face ideal of the Alexander dual of $\Delta_X$.)
The algorithm

**Proposition.** (Jarrah, L., Stigler, Stillman) A subset $F$ of $\{1, \ldots, n\}$ is in $\Delta_X$ if and only if the ideal $\langle x_i \mid i \notin F \rangle$ contains the ideal $M_X$. 
Corollary. To find all possible minimal wiring diagrams, we need to find all minimal subsets of variables $y_1, \ldots, y_m$ such that $M_X$ is contained in $\langle y_1, \ldots, y_m \rangle$. That is, we need to find all minimal primes containing $M_X$. 
Scoring method

Let $F = \{F_1, \ldots, F_t\}$ be the output of the algorithm.
For $s = 1, \ldots, n$, let $Z_s = \#$ sets in $F$ with $s$ elements.
For $i = 1, \ldots, n$, let $W_i(s) = \#$ sets in $F$ of size $s$ which contain $x_i$.

$$S(x_i) := \sum W_i(s) / sZ_s$$
where the sum extends over all $s$ such that $Z_s \neq 0$.

$$T(F_j) := \prod_{x_i \in F_j} S(x_i).$$

Normalization $\Rightarrow$ probability distribution on $F$ of min. var. sets

This scoring method has a bias toward small sets.
Model selection

**Problem:** The model space $f + I$ is WAY TOO BIG

**Solution:** Use “biological theory” to reduce it.
“Biological theory”

• Limit the structure of the coordinate functions $f_i$ to those which are “biologically meaningful.”
  (Characterize special classes computationally.)
• Limit the admissible dynamical properties of models.
  (Identify and computationally characterize classes for which dynamics can be predicted from structure.)
Nested canalyzing functions

Random Boolean network models and the yeast transcriptional network

Stuart Kauffman*, Carsten Peterson†‡, Björn Samuelsson†, and Carl Troein†

*Department of Cell Biology and Physiology, University of New Mexico Health Sciences Center, Albuquerque, NM 87131; and †Complex Systems Division, Department of Theoretical Physics, Lund University, Sölvegatan 14A, 5-223 62 Lund, Sweden

Communicated by Philip W. Anderson, Princeton University, Princeton, NJ, October 6, 2003 (received for review June 30, 2003)

Genetic networks with canalyzing Boolean rules are always stable

Stuart Kauffman*, Carsten Peterson†‡, Björn Samuelsson†‡, and Carl Troein†‡

*Department of Cell Biology and Physiology, Health Sciences Center, University of New Mexico, Albuquerque, NM 87131; and †Complex Systems Division, Department of Theoretical Physics, Lund University, Sölvegatan 14A, 5-223 62 Lund, Sweden
Nested canalyzing functions

Let $\sigma \in S_n$. A Boolean function $f$ in $n$ variables is a \textit{nested canalyzing function} (NCF) in the variable order $x_{\sigma(1)}, \ldots, x_{\sigma(n)}$ with canalyzing input values $a_1, \ldots, a_n$ and canalyzed output values $b_1, \ldots, b_n$, respectively, if

$$f(x) = \begin{cases} 
    b_1 & \text{if } x_{\sigma(1)} = a_1, \\
    b_2 & \text{if } x_{\sigma(1)} \neq a_1 \text{ and } x_{\sigma(2)} = a_2, \\
    b_3 & \text{if } x_{\sigma(1)} \neq a_1 \text{ and } x_{\sigma(2)} \neq a_2 \text{ and } x_{\sigma(3)} = a_3, \\
    \vdots & \vdots \\
    b_n & \text{if } x_{\sigma(1)} \neq a_1 \text{ and } \ldots \text{ and } x_{\sigma(n-1)} \neq a_{n-1} \text{ and } x_{\sigma(n)} = a_n \\
    \overline{b_n} & \text{if } x_{\sigma(1)} \neq a_1 \text{ and } \ldots \text{ and } x_{\sigma(n)} \neq a_n.
\end{cases}$$
A non-canalizing Boolean network

\[
\begin{align*}
  f_1 &= x_4 \\
  f_2 &= x_4 + x_3 \\
  f_3 &= x_2 + x_4 \\
  f_4 &= x_2 + x_1 + x_3
\end{align*}
\]
A nested canalyzing Boolean network

\[
g_1 = x_4 \\
g_2 = x_4 \cdot x_3 \\
g_3 = x_2 \cdot x_4 \\
g_4 = x_2 \cdot x_1 \cdot x_3
\]
Polynomial form of nested canalyzing Boolean functions

**Theorem.** Let $f$ be a function in $\mathcal{R}$. Then

1. $f$ is *canalyzing* in the variable $x_i$, for some $i$, with canalyzing input value $a_i$ and canalyzed output value $b_i$, if and only if

   $$f(x) = (x_i - a_i)g(x_1, x_2, \ldots, x_i, \ldots, x_n) + b_i.$$

2. $f$ is *nested canalyzing* in the order $x_1, x_2, \ldots, x_n$, with canalyzing values $a_i$ and corresponding canalyzed values $b_i$, $1 \leq i \leq n$, if and only if it has the polynomial form

   $$f(x) = (x_1 - a_1)(x_2 - a_2)[(x_3 - a_3)[(x_4 - a_4)[\ldots[(x_{n-1} - a_{n-1})(x_n - a_n)
   + (b_n - b_{n-1})] + (b_{n-1} - b_{n-2})] \ldots] + (b_2 - b_1)] + b_1$$

   or, equivalently,

   $$f(x) = \prod_{i=1}^{n} (x_i - a_i) + \sum_{j=1}^{n-1} \left[ (b_{n-j+1} - b_{n-j}) \prod_{i=1}^{n-j} (x_i - a_i) \right] + b_1.$$
The vector space of Boolean polynomial functions

\[ R = \{ \sum_{S \subseteq [n]} c_S \prod_{i \in S} x_i : c_S \in \mathbb{F}_2 \}. \]

As a vector space over \( \mathbb{F}_2 \), \( R \) is isomorphic to \( \mathbb{F}_2^{2^n} \) via the correspondence

\[ R \ni \sum_{S \subseteq [n]} c_S \prod_{i \in S} x_i \longleftrightarrow (c_\emptyset, \ldots, c_{[n]}) \in \mathbb{F}_2^{2^n}. \]
The variety of nested canalyzing functions

**Corollary.** The point \((c_\emptyset, \ldots, c_{[n]}) \in \mathbb{F}_2^{2^n}\) is the coefficient vector of a nested canalyzing functions in the variable order \(x_1, \ldots, x_n\) if and only if

\[
c_{[n]} = 1 \quad \text{and for } \emptyset \neq S \subseteq [n] \quad c_S = (c_{[r_S]}) \prod_{i \in [r_S] \setminus S} c_{[n] \setminus \{i\}}
\]

The set of all such points is denoted by \(V_{id}^{ncf}\).
Input and output values as functions of the coefficients

**Corollary.** Let $f$ be a Boolean polynomial. If $f$ is nested canonicalizing function in the order $x_1, x_2, \ldots, x_n$, with input values $a_j$ and corresponding output values $b_j, 1 \leq j \leq n$, then

\[ a_j = c[n] \setminus \{j\}, \quad \text{for } 1 \leq j \leq n - 1 \]

\[ b_1 = c_\emptyset + c_1 c[n] \setminus \{1\}, \]

\[ b_{j+1} - b_j = c_{[j+1]} c[n] \setminus \{j+1\} + c[j], \quad \text{for } 1 \leq j < n - 1 \text{ and} \]

\[ b_n - a_n = b_{n-1} + c[n-1]. \]
Corollary.
The ideal of relations defining the class of nested canalyzing Boolean functions on $n$ variables forms an affine toric variety over the algebraic closure of $F_2$. The irreducible components correspond to the functions that are nested canalyzing with respect to a given variable ordering.

(joint work with Jarrah, Raposa)
Dynamics from structure

**Theorem.** Let $f = (f_1, \ldots, f_n) : k^n \rightarrow k^n$ be a monomial system.

1. If $k = F_2$, then $f$ is a fixed point system if and only if every strongly connected component of the dependency graph of $f$ has loop number 1. (Colón-Reyes, L., Pareigis)

2. The case for general $k$ can be reduced the Boolean + linear case. (Colón-Reyes, Jarrah, L., Sturmfels)
Questions

• What are good classes of functions from a biological and/or mathematical point of view?
• What extra mathematical structure is needed to make progress?
• How does the nature of the observed data points affect the structure of $f+I$ and $M_X$?
Modeling and Simulation of Biological Networks

Symposia in Pure and Applied Math, AMS

in press

articles by Allman-Rhodes, Pachter, Stigler, ….
Special year 2008-09 at SAMSI:

“Algebraic methods in biology and statistics”

(subject to final approval)