What is Algebraic Biology?

Matthew Macauley

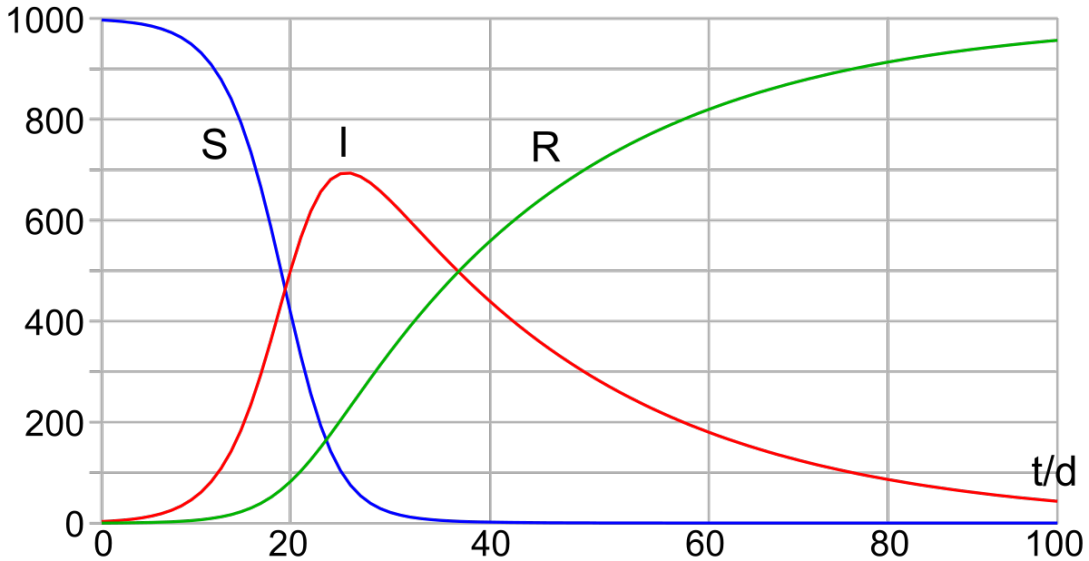
Department of Mathematical Sciences
Clemson University
http://www.math.clemson.edu/~macaule/

Algebraic Biology

What does algebra have to do with biology?

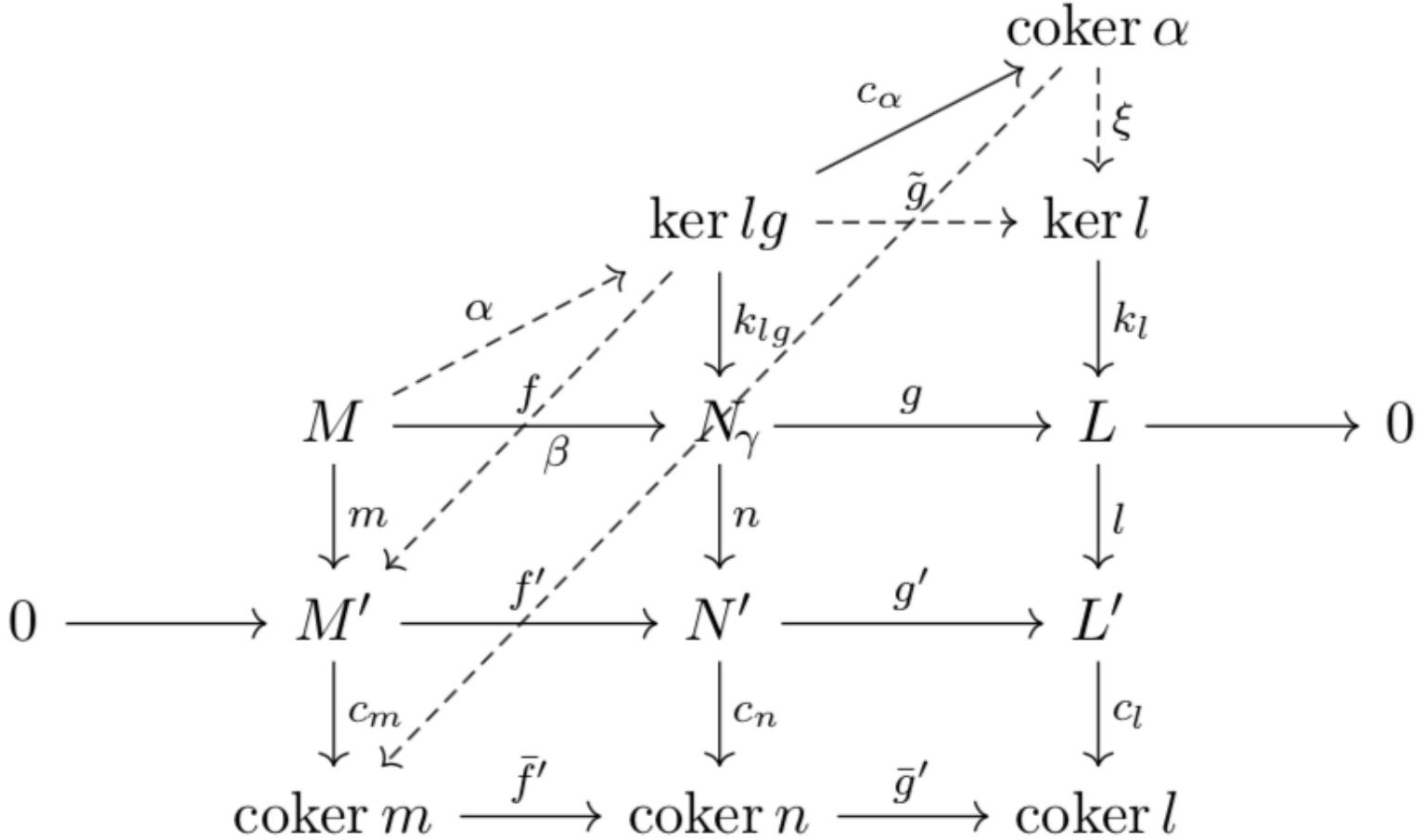
Usually, when we think of mathematical biology, we think of models such as this:

$$\frac{dS}{dt} = -\alpha SI$$
$$\frac{dI}{dt} = \alpha SI - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$



What does algebra have to do with biology?

Whereas algebra might remind us more of this:

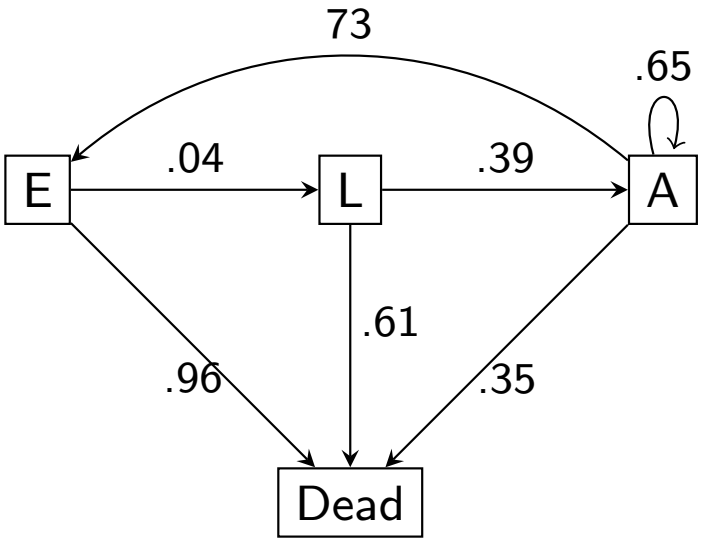


How could these two topics possibly be related??

What does algebra have to do with biology?

We all know that linear algebra is fundamental to mathematical biology.

Consider the following example, of a structured population of Eggs, Larvae, and Adults.



$$\begin{bmatrix} 0 & 0 & 73 \\ .04 & 0 & 0 \\ .65 & .39 & 0 \end{bmatrix} \begin{bmatrix} E_t \\ L_t \\ A_t \end{bmatrix} = \begin{bmatrix} E_{t+1} \\ L_{t+1} \\ A_{t+1} \end{bmatrix}.$$

This is one example of many, of how linear differential or difference equations can model natural phenomena.

What does algebra have to do with biology?

Linear algebra also arises when approximating non-linear models, a process called linearization.

For example, consider the following Lotka-Volterra equations that model two competing species:

$$P' = P(1 - P - Q)$$
$$Q' = Q(.75 - Q - .5P)$$

Since $(1, 0)$ is a steady-state, we can change variables $(X, Y) = (P - 1, Q - 0)$, and get the system

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & .25 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} - \begin{bmatrix} X^2 + XY \\ .5XY + Y^2 \end{bmatrix}.$$

For $(X, Y) \approx (1, 0)$, the non-linear terms are negligible. The *linearized* system is thus

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} \approx \begin{bmatrix} -1 & -1 \\ 0 & .25 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}.$$

What does algebra have to do with biology?

Linear algebra, the study of linear polynomials and their solutions, is a fundamental pillar of mathematical biology.

Analyzing nonlinear polynomials and their solutions is much more complex.

It involves fields such as algebraic geometry and computational algebra.

Though these themes are not as ubiquitous in biology as linear algebra is, they arise in a number of biological problems.

Algebraic Biology is the subfield that encompasses these problems, and the new mathematics that they spawn.

In the rest of this lecture, we'll see four examples of biological problems where nonlinear algebra arises:

1. Biochemical reaction networks
2. Boolean models of molecular networks
3. Algebraic statistics and phylogenetics
4. Place fields in neuroscience

Biochemical reaction networks

Consider a simple biochemical reaction, where A, B, and C are molecular species:

$$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C, \qquad A \xrightarrow{k_3} 2B.$$

The constants k_1, k_2, and k_3 represent reaction rates.

Let $x_1(t)$, $x_2(t)$, and $x_3(t)$ denote concentrations of A, B, and C. The assumption of the laws of mass-action kinetics leads to the following system of ODEs:

$$\begin{aligned} x_1' &= -k_1 x_1 x_2 - k_3 x_1 + k_2 x_3 \\ x_2' &= -k_1 x_1 x_2 + k_2 x_3 + 2k_3 x_1 \\ x_3' &= k_1 x_1 x_2 - k_2 x_3. \end{aligned}$$

To find the steady-states, set each $x_i' = 0$ and solve the system.

Biologically, we only care about solutions in the non-negative orthant of $\mathbb{R}^3$. However, polynomials are easier to study over $\mathbb{C}$.

In the language of algebraic geometry, for each fixed choice of parameters, the solutions to the system above form an algebraic variety in $\mathbb{C}^3$.

This can be found by computing a Gröbner basis of the ideal

$$I = \langle -k_1 x_1 x_2 - k_3 x_1 + k_2 x_3,\ -k_1 x_1 x_2 + k_2 x_3 + 2k_3 x_1,\ k_1 x_1 x_2 - k_2 x_3 \rangle.$$

Boolean models of molecular networks

The following is a Boolean model of the lactose (*lac*) operon in *E. coli*:

$$\begin{aligned} x_1(t+1) &= \neg G_e \wedge (x_3(t) \vee L_e) \\ x_2(t+1) &= x_1(t) \\ x_3(t+1) &= \neg G_e \wedge [(L_e \wedge x_2(t)) \vee (x_3(t) \wedge \neg x_2(t))]. \end{aligned}$$

Time is discretized, and $x_1(t)$, $x_2(t)$, and $x_3(t)$ represent mRNA, translated proteins, and lactose.

The steady-states are found by setting each $x_i(t+1) = x_i(t)$ and solving the resulting system.

In polynomial form, this system is

$$\begin{aligned} (1+G_e)(x_3L_e + x_3 + L_3) + x_1 &= 0 \\ x_1 + x_2 &= 0 \\ (1+G_e)(x_2L_e + x_3(1+x_2)) + x_3 &= 0. \end{aligned}$$

The soluions can be found by computing a Gröbner basis of the ideal

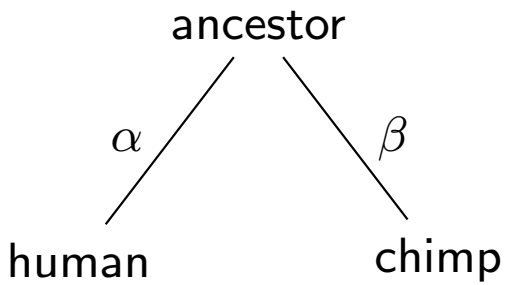
$$I = \left\langle (1+G_e)(x_3L_e + x_3 + L_e) + x_1,\ x_1 + x_2,\ (1+G_e)(x_2L_e + x_3(1+x_2)) + x_3 \right\rangle.$$

Phylogenetics

Consider a simple evolutionary relationship of two species and their most common ancestor.

Fix a particular base in the genome at a site that all three species share in a mutual alignment.

Under the Jukes-Cantor model of evolution, the probability of a mutation at that site is a constant.



It is straightforward to compute the probability that (human,chimp)=(A,C):

$$P(AC) = P\left(\begin{array}{ccc} & A & \\ A & & C \end{array}\right) + P\left(\begin{array}{ccc} & G & \\ A & & C \end{array}\right) + P\left(\begin{array}{ccc} & C & \\ A & & C \end{array}\right) + P\left(\begin{array}{ccc} & T & \\ A & & C \end{array}\right)$$

$$= \frac{1}{4}(1-3\alpha)\beta + \frac{1}{4}\alpha\beta + \frac{1}{4}\alpha(1-3\beta) + \frac{1}{4}\alpha\beta = \frac{1}{4}(\alpha+\beta-\alpha\beta).$$

Phylogenetics

Similarly, $P(AA) = \frac{1}{4}(1-3\alpha)(1-3\beta) + \frac{3}{4}\alpha\beta = 3\alpha\beta + \frac{1}{4}(1-3\alpha-3\beta)$.

The space of possible probabilities can be described by a mapping

$$\varphi\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^{16}, \qquad \varphi\colon (\alpha,\beta) \longmapsto \big(P(AA), P(AC), \ldots, P(TT)\big).$$

For an n-leaf tree with $m = 2n-2$ edges, we get a map $\varphi\colon \mathbb{R}^m \to \mathbb{R}^{4^n}$.

The intersection of $\mathrm{Im}(\varphi)$, with the $d = 4^n - 1$ dimensional simplex Δ_d is the *phylogenetic model*, $\mathcal{M}_T \subseteq \mathbb{R}^{4^n}$.

The polynomials that vanish on $\mathcal{M}_T$ is called the ideal of phylogenetic invariants,

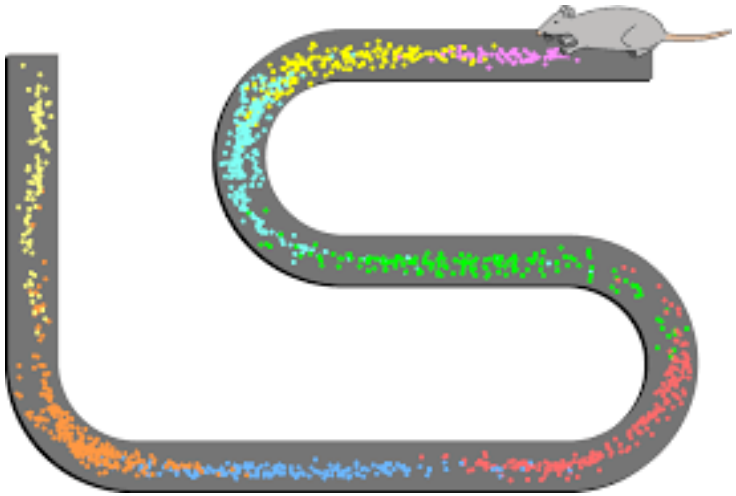
$$I_T = I_T(\mathcal{M}_T) = \big\{ f \in \mathbb{R}[x_1, \ldots, x_{4^n}] \mid f(p) = 0, \quad \text{for all } p \in \mathcal{M}_T \big\}.$$

The points that vanish on all polynomials in the ideal I_T is called the phylogenetic variety of T:

$$V_T = V_T(I_T) = \big\{ p \in \mathbb{R}^{4^n} \mid f(p) = 0, \quad \text{for all } f \in I_T \big\}.$$

Place fields in neuroscience

Experiments have shown that neurons called place cells fire based on an animal's location.



As an animal moves around, different subsets of neurons fire. The region that causes a specific neuron to fire is its place field.

We can encode which neurons fire with a binary string. For example, $\mathbf{c} = 10100$ means neurons 1 and 3 fire, and neurons 2, 4, and 5 are silent.

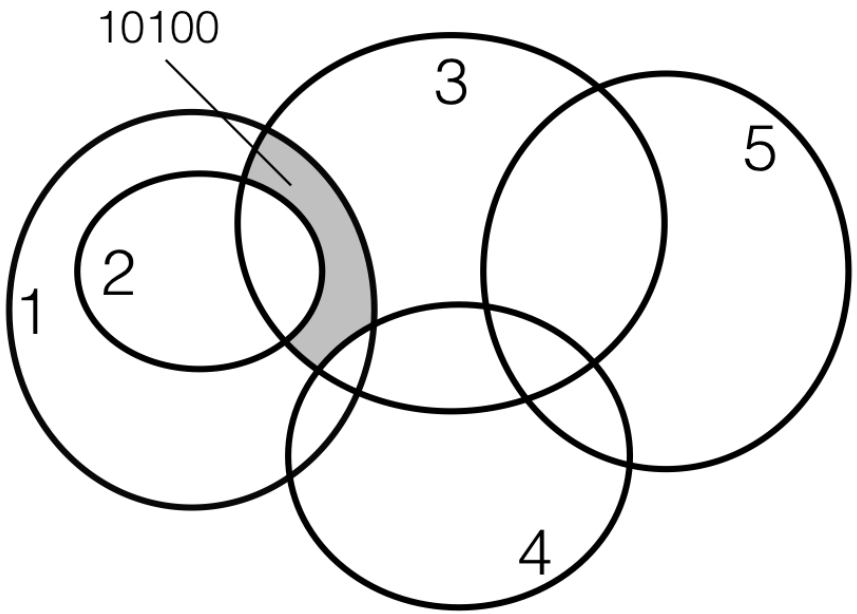
Place fields in neuroscience

Motivating question

Given a collection of binary strings called a neural code, reconstruct the place fields.

For example, how would you construction place fields $\mathcal{U} = \{U_1, U_2, U_3, U_4, U_5\}$ that realize the code

$$\mathcal{C} = \{00000, 10000, 11000, 10100, 11100, 10010, 10110,$$
$$00100, 00110, 00101, 00111, 00010, 00011, 00001\}?$$



Place fields in neuroscience

Another interesting question

Given a neural code, can it be realized by a collection of open convex place fields?

For example, the code

$$\mathcal{C} = \{000, 100, 010, 101, 110, 011\}$$

cannot be realized by open convex place fields.

Many of these questions can be approached algebraically. Every code $\mathcal{C}$ has a *vanishing ideal*,

$$I_{\mathcal{C}} = \big\{ f \in \mathbb{F}_2[x_1, \ldots, x_n] \mid f(\mathbf{c}) = 0 \text{ for all } \mathbf{c} \in \mathcal{C} \big\}.$$

A related object is the neural ideal, which is defined by the *characteristic polynomials* of the non-code words:

$$J_{\mathcal{C}} = \big\langle \chi_{\mathbf{n}}(\mathbf{x}) \mid \mathbf{n} \notin \mathcal{C} \big\rangle, \qquad \text{where} \quad \chi_{\mathbf{n}}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} = \mathbf{n} \\ 0 & \mathbf{x} \neq \mathbf{n}. \end{cases}$$