Packages for biological applications: PhylogeneticTrees.m2 and ReactionNetworks.m2

Elizabeth Gross (SJSU)

July 29, 2017

Packages for Biological Applications

`PhylogeneticTrees.m2` Hector Baños, Nathaniel Bushek, Ruth Davidson, Elizabeth Gross, Pamela Harris, Robert Krone, Colby Long, Allen Stewart, Robert Walker.

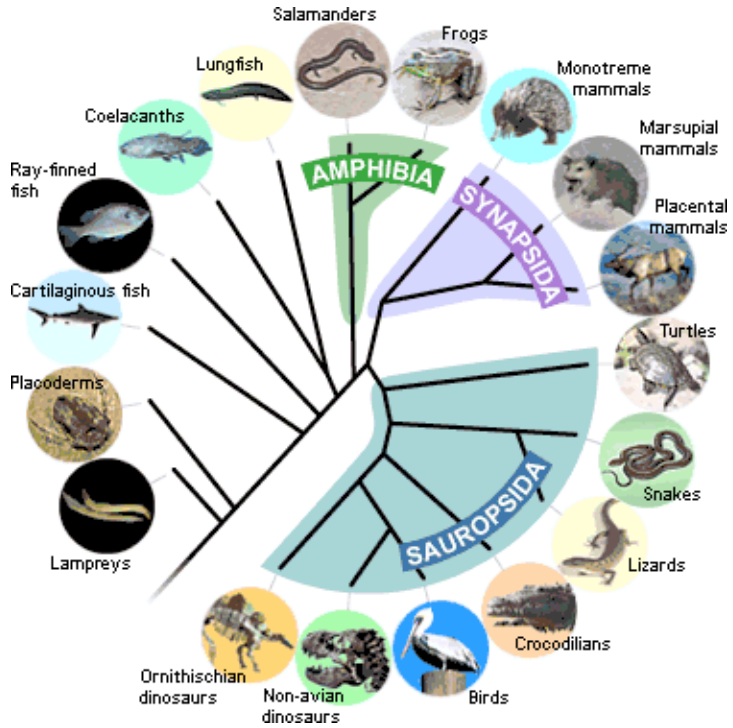
`ReactionNetworks.m2` Timothy Duff, Cvetelina Hill, Kisun Lee, Anton Leykin.

PhylogeneticTrees.m2

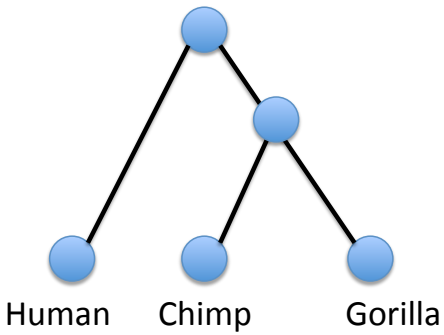
Inferring phylogenetic trees

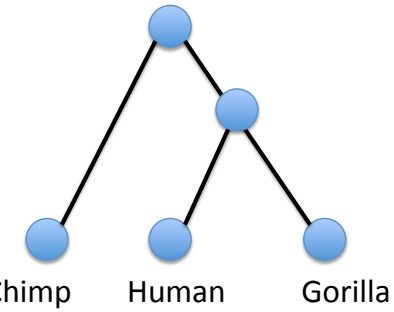
Problem:
Given aligned DNA sequences from a collection of species, find the tree that best describes the species' ancestral history.

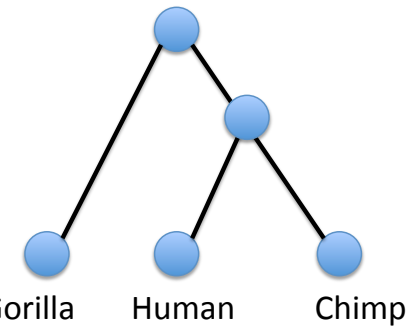
Human : ...ACCGTGCAACGTGAACGA...
Chimp : ...ACCTTGCAAGGTAAACGA...
Gorilla : ...ACCGTGCAACGTAAACTA...



Possible Trees:







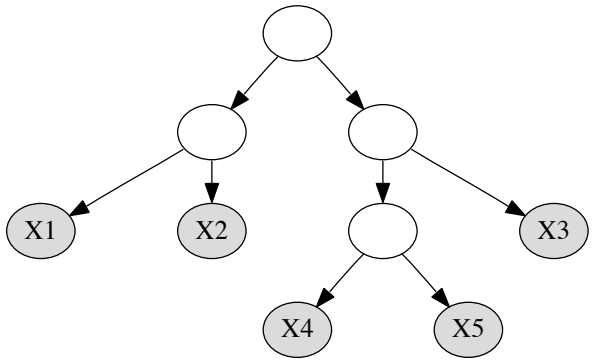
Tree-based Markov models

- Assumes evolution proceeds along a n-leaf tree according to a Markov process.

- Assumes site independence.

- Data are the observed frequencies of all n-tuples of DNA bases.

```
...ACCGTGCAACGTGAACGA...
...ACCTTGCAAGGTAAACGA...
...ACCGTGCAACGTAAACTA...
```



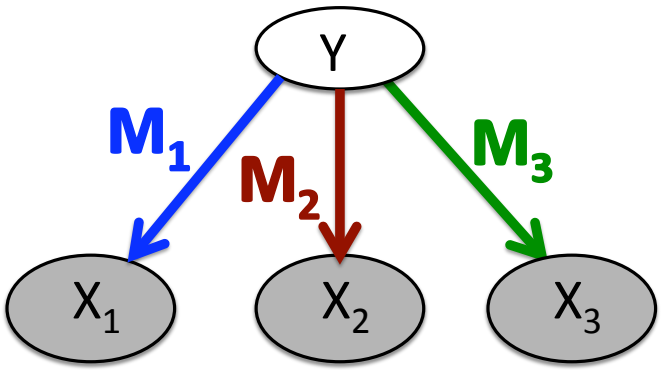
Gray nodes:
extant species (observable)

White nodes:
extinct species (hidden)

Group-based Markov models

Parameters: A tree T and transition matrices for each edge.

Example: 4-state group-based Markov model (K3P) on the claw tree $K_{1,3}$



$$\mathbf{M_1} = \begin{bmatrix} P_{A|A} & P_{C|A} & P_{G|A} & P_{T|A} \\ P_{A|C} & P_{C|C} & P_{G|C} & P_{T|C} \\ P_{A|G} & P_{C|G} & P_{G|G} & P_{T|G} \\ P_{A|T} & P_{C|T} & P_{G|T} & P_{T|T} \end{bmatrix}$$

where $P_{i|j} = P(X_1 = i \mid Y = j)$.

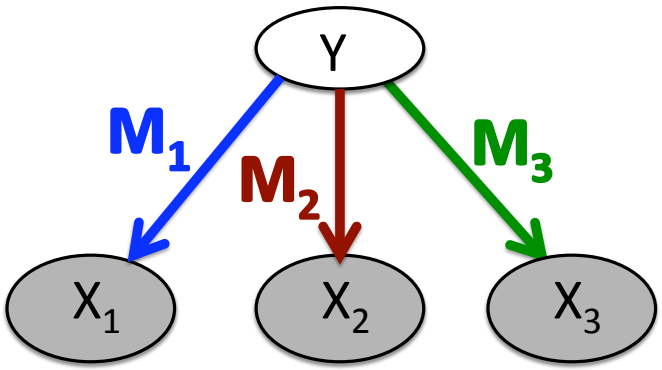
$X_1, X_2, X_3 \in \{A, C, G, T\}$ are random variables and $\{A, C, G, T\}$ is viewed as the group $\mathbb{Z}_2 \otimes \mathbb{Z}_2$.

$Y \in \{A, C, G, T\}$ is a hidden (latent) random variable with distribution $(\pi_A, \pi_C, \pi_G, \pi_T)$, e.g. $P(Y = A) = \pi_A$.

Group-based Markov models

Parameters: A tree T and transition matrices for each edge.

Example: 4-state group-based Markov model (K3P) on the claw tree $K_{1,3}$



$$\mathbf{M_1}, \mathbf{M_2}, \mathbf{M_3} = \begin{bmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha \end{bmatrix}$$

$X_1, X_2, X_3 \in \{A, C, G, T\}$ are random variables and $\{A, C, G, T\}$ is viewed as the group $\mathbb{Z}_2 \otimes \mathbb{Z}_2$.

$Y \in \{A, C, G, T\}$ is a hidden (latent) random variable with distribution $(\pi_A, \pi_C, \pi_G, \pi_T)$, e.g. $P(Y = A) = \pi_A$.

Transition matrices

Cavender-Farris-Neyman (CFN)

$$\begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}$$

Jukes-Cantor (JC)

$$\begin{pmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{pmatrix}$$

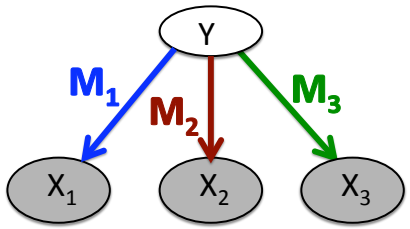
Kimura 2-parameter (K2P)

$$\begin{pmatrix} \alpha & \beta & \gamma & \gamma \\ \beta & \alpha & \gamma & \gamma \\ \gamma & \gamma & \alpha & \beta \\ \gamma & \gamma & \beta & \alpha \end{pmatrix}$$

Kimura 3-parameter (K3P)

$$\begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha \end{pmatrix}$$

Models, Ideals, and Varieties



The parameterization of the model $\mathcal{M}_T$ (K3P) is

$$\phi_T : \mathbb{R}^4 \times \mathbb{R}^4 \times \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R}^{4\times4\times4}$$

$$(\pi, \mathbf{M_1}, \mathbf{M_2}, \mathbf{M_3}) \mapsto \sum_{i=1}^{4} \pi_i \mathbf{M_{1}}_i \otimes \mathbf{M_{2}}_i \otimes \mathbf{M_{3}}_i$$

Image in $\mathbb{R}^{4\times4\times4}$ of a point in the parameter space is a probability table p whose jklth entry is the joint probability that $X_1 = j, X_2 = k$, and $X_3 = l$.

$$p_{jkl} = \sum_{i=1}^{4} \pi_i \mathbf{M_{1}}_{ij} \mathbf{M_{2}}_{ik} \mathbf{M_{3}}_{il}.$$

The **ideal** *associated to* $\mathcal{M}_T$ *is*

$$\mathcal{I}_T = \{f \in \mathbb{C}[p_{jkl} \ : \ j,k,l \in \{A,C,G,T\}] \ : f(p) = 0 \text{ for all } p \in \mathcal{M}_T\}$$

The **variety** *associated to* $\mathcal{M}_T$ *is*

$$\mathcal{V}_T = \{p \in \mathbb{C}^{4\times4\times4} \ : \ f(p) = 0 \text{ for all } f \in \mathcal{I}_T\} = \overline{Im\ \phi_T} = \overline{\mathcal{M}_T}.$$

Group-based models correspond to toric varieties

Theorem (Hendy-Penny 1993, Evans-Speed 1993)

In the Fourier coordinates, a group-based model is parametrized by monomial functions in terms of the Fourier parameters. (See Sturmfels-Sullivant 2005 for detailed description)

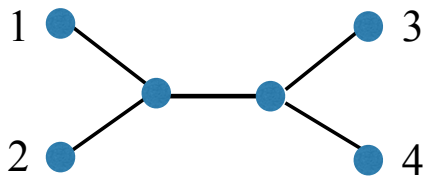
- G: $\mathbb{Z}_2$ or $\mathbb{Z}_2 \times \mathbb{Z}_2$
- T: n taxon tree.
- $\Sigma(T)$: set of splits of T.
- For split $A|B \in \Sigma(T)$, associate a set of parameters: $a_g^{A|B}$ where $g \in G$.

The toric parameterization for the model is:

$$q_{g_1,\ldots,g_n} = \begin{cases} \prod_{A|B \in \Sigma(T)} a^{A|B}_{\sum_{i \in A} g_i} & \text{if } \sum_{i=1}^n g_i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Example

Kimura 3-parameter model



$$\Sigma(T) = \{1|234, 2|134, 3|124, 4|123, 12|34\}$$

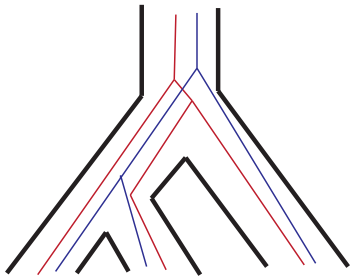
Parameterization:

$$q_{g_1 g_2 g_3 g_4} = a_{g_1}^{1|234} a_{g_2}^{2|134} a_{g_3}^{3|124} a_{g_4}^{4|123} a_{g_1+g_2}^{12|34}$$

Example:

$$q_{ACGT} = a_A^{1|234} a_C^{2|134} a_G^{3|124} a_T^{4|123} a_C^{12|34}$$

Mixture models



Due to biological mechanisms, such as incomplete lineage sorting or horizontal gene transfer, sometimes we want to consider the **mixture** of two tree models.

- T_1, T_2: n leaf trees
- $\mathcal{M}_{T_1}$, $\mathcal{M}_{T_2}$: tree-based models
- ϕ_{T_1}, ϕ_{T_2}: parameterization maps of $\mathcal{M}_{T_1}$ and $\mathcal{M}_{T_2}$
- α: the *mixing parameter*

The parameterization of the mixture model $\mathcal{M}_{T_1,T_2}$ is

$$\psi_{T_1,T_2} : \Theta_{T_1} \times \Theta_{T_2} \times [0,1] \to \Delta^{4^n-1} \subseteq \mathbb{R}^{4^n}$$
$$(\theta_1, \theta_2, \alpha) \mapsto \alpha\phi_{T_1}(\theta_1) + (1-\alpha)\phi_{T_2}(\theta_2)$$

The corresponding variety of $\mathcal{M}_{T_1,T_2}$ is a **join** variety.

$$\mathcal{V}_{T_1,T_2} = \overline{\mathcal{M}_{T_1,T_2}} = \overline{Im\ \psi_{T_1,T_2}} = Join(V_{T_1}, V_{T_2})$$

Open Problems for mixture models

- **Determine invariants for mixture models** These invariants can be used for model selection and also to prove theoretical results regarding identifiability.
- **Identifiability** Determine when

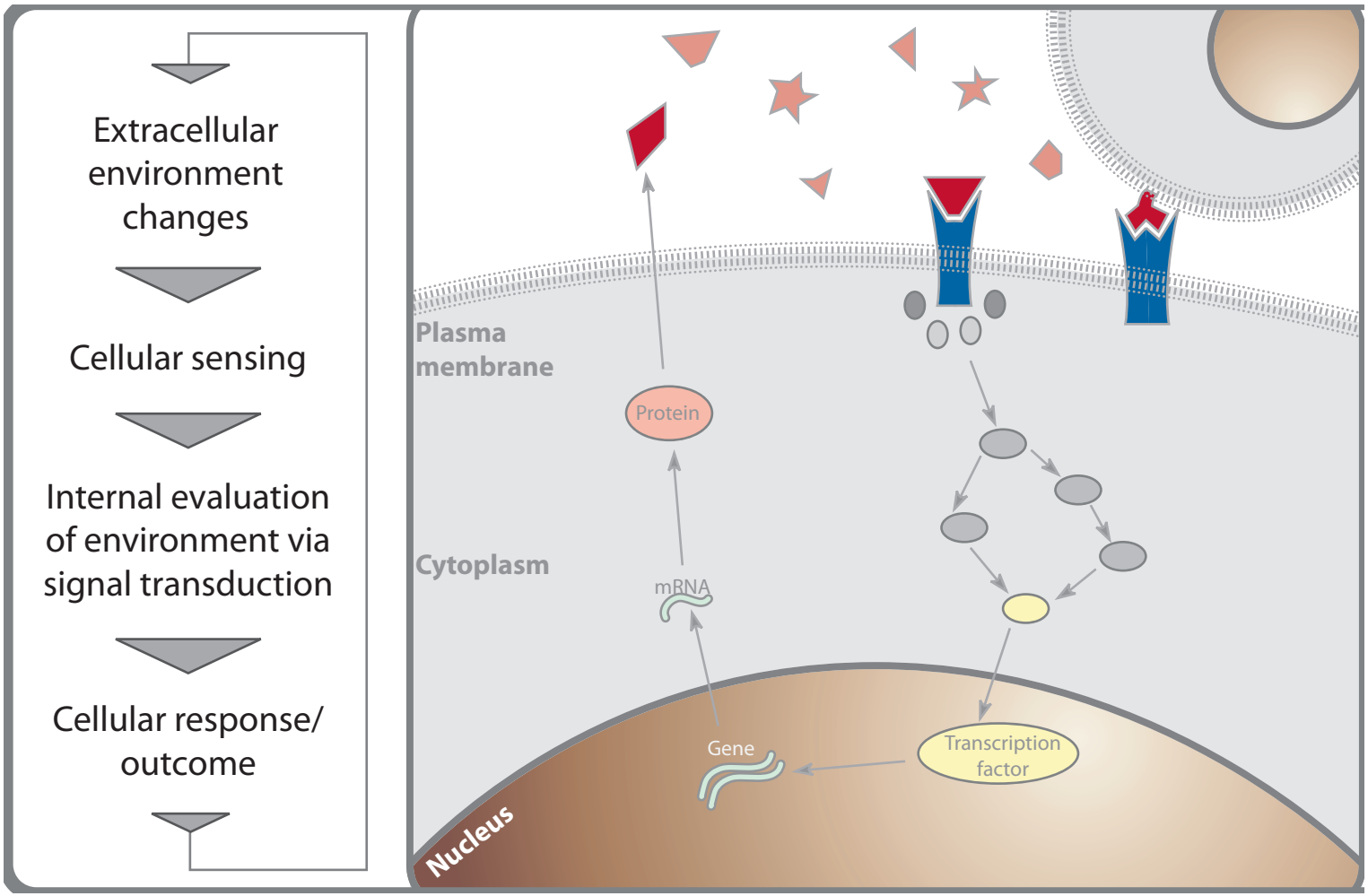
$$\mathcal{V}_{\mathcal{T}_1,\mathcal{T}_2} \subseteq \mathcal{V}_{\mathcal{T}_3,\mathcal{T}_4}$$

To establish identifiability, one usually needs to know

1. The dimension of $\mathcal{V}_{\mathcal{T}_1,\mathcal{T}_2}$ and $\mathcal{V}_{\mathcal{T}_3,\mathcal{T}_4}$ (current work with Hector Baños, Nathaniel Bushek, Ruth Davidson, Elizabeth Gross, Pamela Harris, Robert Krone, Colby Long, Allen Stewart, and Robert Walker).
2. Some invariants of $\mathcal{M}_{\mathcal{T}_1,\mathcal{T}_2}$.

ReactionNetworks.m2

How do cells make decisions?



Chemical Reaction Network Theory

A chemical reaction network is a given by a triple $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ of finite sets.

- **Species**, $\mathcal{S} = \{S_1, \ldots, S_d\}$: molecules undergoing a series of chemical reactions.
- **Complexes**, $\mathcal{C} = \{C_1, \ldots, C_n\}$: linear combinations of the species representing those used and produced in each reaction (i.e. *reactants and products*).
- **Reactions**, $\mathcal{R} = \{y_j \to y_j'\}$: directed graph with the complexes as vertices, $y_j, y_j' \in \mathcal{C}$

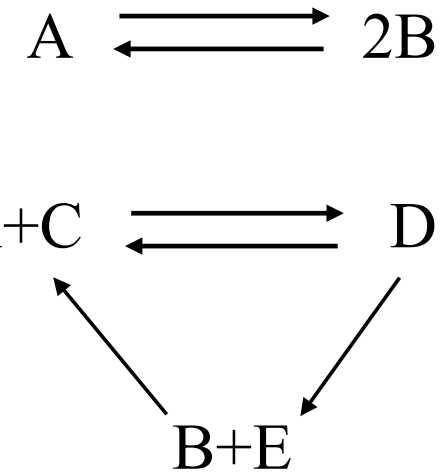
Example

$$A + B \to 2B$$

$$B \to A$$

$$\mathcal{S} = \{A, B\}, \quad \mathcal{C} = \{A + B, 2B, B, A\}, \quad \mathcal{R} = \{A + B \to 2B, B \to A\}$$

Mass action kinetics



$$\mathcal{S} = \{A, B, C, D, E\}$$
$$\mathcal{C} = \{A, 2B, A + C, D, B + E\}$$

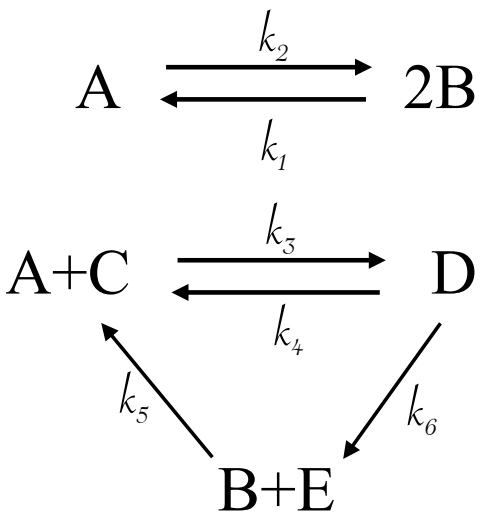
We will work in the deterministic setting with the assumption of mass action kinetics.

Definition

Mass-action kinetics: rate of reaction is proportional to the product of the concentrations of the species.

We call the constant of proportionality the rate constant.

Mass action kinetics



$$\mathcal{S} = \{A, B, C, D, E\}$$
$$\mathcal{C} = \{A, 2B, A + C, D, B + E\}$$

We will work in the deterministic setting with the assumption of mass action kinetics.

Definition

Mass-action kinetics: rate of reaction is proportional to the product of the concentrations of the species.

We call the constant of proportionality the rate constant.

Polynomial dynamical systems

The assumption of mass-action kinetics leads to polynomial dynamical systems that can be read off from the network.

$$A + B \xrightarrow{k_1} 2B$$
$$B \xrightarrow{k_2} A$$

Let x_A and x_B denote the concentrations of the species A and B.

Each complex corresponds to a monomial:

$$A + B: \ x_A x_B, \quad 2B: \ x_B^2, \quad A: \ x_A, \quad B: \ x_B,$$

$$\frac{d}{dt} x_A = ?$$

$$\frac{d}{dt} x_B = ?$$

Polynomial dynamical systems

The assumption of mass-action kinetics leads to polynomial dynamical systems that can be read off from the network.

$$A + B \xrightarrow{k_1} 2B$$
$$B \xrightarrow{k_2} A$$

Let x_A and x_B denote the concentrations of the species A and B.

Each complex corresponds to a monomial:

$$A + B: \ x_A x_B, \quad 2B: \ x_B^2, \quad A: \ x_A, \quad B: \ x_B,$$

$$\frac{d}{dt} x_A = -k_1 x_A x_B$$

$$\frac{d}{dt} x_B = ?$$

Polynomial dynamical systems

The assumption of mass-action kinetics leads to polynomial dynamical systems that can be read off from the network.

$$A + B \xrightarrow{k_1} 2B$$
$$B \xrightarrow{k_2} A$$

Let x_A and x_B denote the concentrations of the species A and B.

Each complex corresponds to a monomial:

$$A + B: \ x_A x_B, \quad 2B: \ x_B^2, \quad A: \ x_A, \quad B: \ x_B,$$

$$\frac{d}{dt} x_A = -k_1 x_A x_B + k_2 x_B$$

$$\frac{d}{dt} x_B = ?$$

Polynomial dynamical systems

The assumption of mass-action kinetics leads to polynomial dynamical systems that can be read off from the network.

$$A + B \xrightarrow{k_1} 2B$$
$$B \xrightarrow{k_2} A$$

Let x_A and x_B denote the concentrations of the species A and B.

Each complex corresponds to a monomial:

$$A + B: \ x_A x_B, \quad 2B: \ x_B^2, \quad A: \ x_A, \quad B: \ x_B,$$

$$\frac{d}{dt} x_A = -k_1 x_A x_B + k_2 x_B$$

$$\frac{d}{dt} x_B = k_1 x_A x_B$$

Polynomial dynamical systems

The assumption of mass-action kinetics leads to polynomial dynamical systems that can be read off from the network.

$$A + B \xrightarrow{k_1} 2B$$
$$B \xrightarrow{k_2} A$$

Let x_A and x_B denote the concentrations of the species A and B.

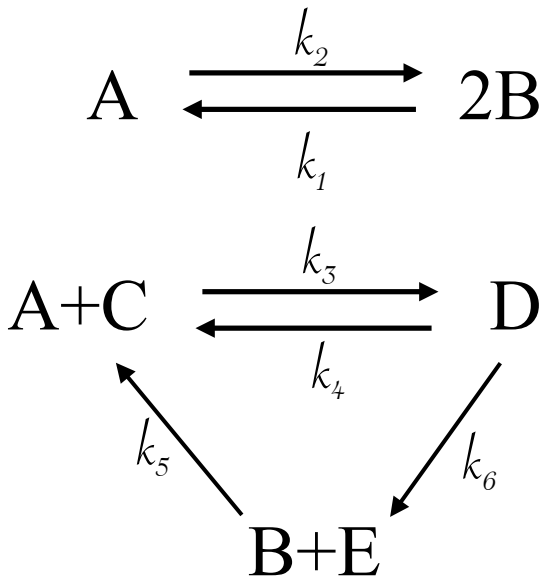
Each complex corresponds to a monomial:

$$A + B: \ x_A x_B, \quad 2B: \ x_B^2, \quad A: \ x_A, \quad B: \ x_B,$$

$$\frac{d}{dt} x_A = -k_1 x_A x_B + k_2 x_B$$

$$\frac{d}{dt} x_B = k_1 x_A x_B - k_2 x_B$$

A larger example



$$\dot{x_A} = k_1 x_B^2 - k_2 x_A + k_3 x_D - k_4 x_A x_C + k_5 x_B x_E$$
$$\dot{x_B} = -2k_1 x_B^2 + 2k_2 x_A - k_5 x_B x_E + k_6 x_D$$
$$\dot{x_C} = k_3 x_D - k_4 x_A x_C + k_5 x_B x_E$$
$$\dot{x_D} = -k_3 x_D + k_4 x_A x_C - k_6 x_D$$
$$\dot{x_E} = -k_5 x_B x_E + k_6 x_D$$

An even larger example

Shuttle model for Wnt signaling pathway

MacLean, Rosen, Byrne, Harrington 2015

$$x_1 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} x_2$$

$$x_2 + x_4 \underset{k_4}{\overset{k_3}{\rightleftharpoons}} x_{14} \xrightarrow{k_5} x_2 + x_5$$

$$x_5 + x_8 \underset{k_7}{\overset{k_6}{\rightleftharpoons}} x_{16} \xrightarrow{k_8} x_4 + x_8$$

$$x_4 + x_{10} \underset{k_{10}}{\overset{k_9}{\rightleftharpoons}} x_{18} \xrightarrow{k_{11}} x_4 + \emptyset$$

$$\emptyset \xrightarrow{k_{12}} x_{10}$$

$$x_{10} \xrightarrow{k_{13}} \emptyset$$

$$x_3 + x_6 \underset{k_{15}}{\overset{k_{14}}{\rightleftharpoons}} x_{15} \xrightarrow{k_{16}} x_3 + x_7$$

$$x_7 + x_9 \underset{k_{18}}{\overset{k_{17}}{\rightleftharpoons}} x_{17} \xrightarrow{k_{19}} x_6 + x_9$$

$$x_6 + x_{11} \underset{k_{21}}{\overset{k_{20}}{\rightleftharpoons}} x_{19} \xrightarrow{k_{22}} x_6 + \emptyset$$

$$x_{11} \xrightarrow{k_{23}} \emptyset$$

$$x_{11} + x_{12} \underset{k_{25}}{\overset{k_{24}}{\rightleftharpoons}} x_{13}$$

$$x_2 \underset{k_{27}}{\overset{k_{26}}{\rightleftharpoons}} x_3$$

$$x_5 \underset{k_{29}}{\overset{k_{28}}{\rightleftharpoons}} x_7$$

$$x_{10} \underset{k_{31}}{\overset{k_{30}}{\rightleftharpoons}} x_{11}$$

Shuttle model for Wnt signaling pathway

$$x_1 \underset{k_2}{\overset{k_1}{\rightleftarrows}} x_2$$

$$x_2 + x_4 \underset{k_4}{\overset{k_3}{\rightleftarrows}} x_{14} \xrightarrow{k_5} x_2 + x_5$$

$$x_5 + x_8 \underset{k_7}{\overset{k_6}{\rightleftarrows}} x_{16} \xrightarrow{k_8} x_4 + x_8$$

$$x_4 + x_{10} \underset{k_{10}}{\overset{k_9}{\rightleftarrows}} x_{18} \xrightarrow{k_{11}} x_4 + \emptyset$$

$$\emptyset \xrightarrow{k_{12}} x_{10}$$

$$x_{10} \xrightarrow{k_{13}} \emptyset$$

$$x_3 + x_6 \underset{k_{15}}{\overset{k_{14}}{\rightleftarrows}} x_{15} \xrightarrow{k_{16}} x_3 + x_7$$

$$x_7 + x_9 \underset{k_{18}}{\overset{k_{17}}{\rightleftarrows}} x_{17} \xrightarrow{k_{19}} x_6 + x_9$$

$$x_6 + x_{11} \underset{k_{21}}{\overset{k_{20}}{\rightleftarrows}} x_{19} \xrightarrow{k_{22}} x_6 + \emptyset$$

$$x_{11} \xrightarrow{k_{23}} \emptyset$$

$$x_{11} + x_{12} \underset{k_{25}}{\overset{k_{24}}{\rightleftarrows}} x_{13}$$

$$x_2 \underset{k_{27}}{\overset{k_{26}}{\rightleftarrows}} x_3$$

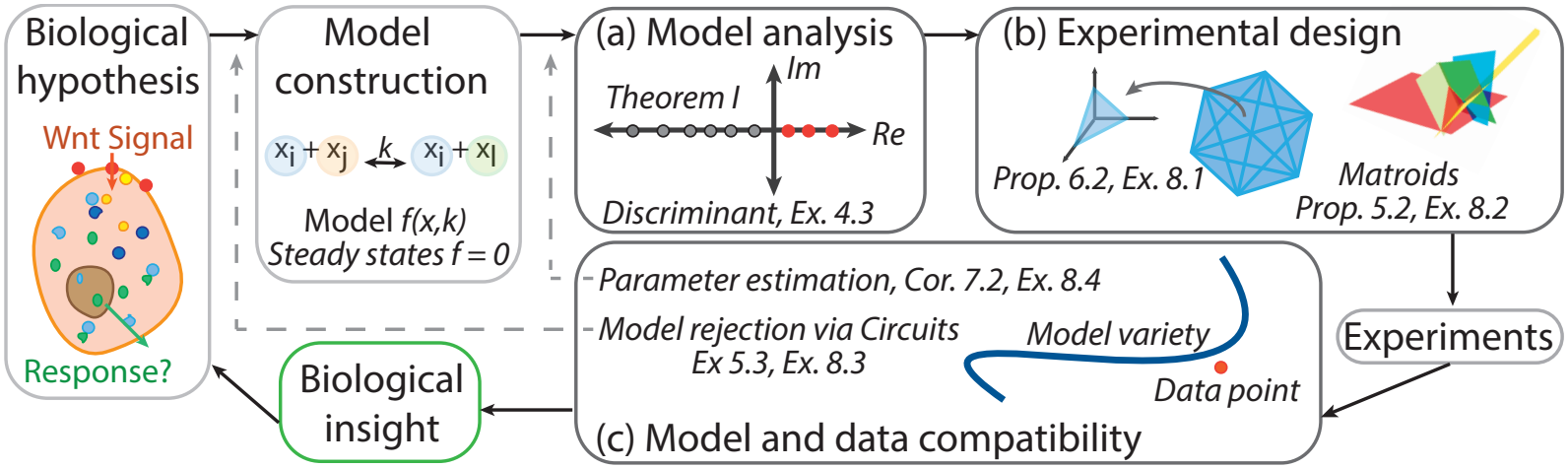
$$x_5 \underset{k_{29}}{\overset{k_{28}}{\rightleftarrows}} x_7$$

$$x_{10} \underset{k_{31}}{\overset{k_{30}}{\rightleftarrows}} x_{11}$$

$$\begin{aligned}
\dot{x}_1 &= -k_1x_1 + k_2x_2 \\
\dot{x}_2 &= k_1x_1 - (k_2 + k_{26})x_2 + k_{27}x_3 - k_3x_2x_4 + (k_4 + k_5)x_{14} \\
\dot{x}_3 &= k_{26}x_2 - k_{27}x_3 - k_{14}x_3x_6 + (k_{15} + k_{16})x_{15} \\
\dot{x}_4 &= -k_3x_2x_4 - k_9x_4x_{10} + k_4x_{14} + k_8x_{16} + (k_{10} + k_{11})x_{18} \\
\dot{x}_5 &= -k_{28}x_5 + k_{29}x_7 - k_6x_5x_8 + k_5x_{14} + k_7x_{16} \\
\dot{x}_6 &= -k_{14}x_3x_6 - k_{20}x_6x_{11} + k_{15}x_{15} + k_{19}x_{17} + (k_{21} + k_{22})x_{19} \\
\dot{x}_7 &= k_{28}x_5 - k_{29}x_7 - k_{17}x_7x_9 + k_{16}x_{15} + k_{18}x_{17} \\
\dot{x}_8 = -\dot{x}_{16} &= -k_6x_5x_8 + (k_7 + k_8)x_{16} \\
\dot{x}_9 = -\dot{x}_{17} &= -k_{17}x_7x_9 + (k_{18} + k_{19})x_{17} \\
\dot{x}_{10} &= k_{12} - (k_{13} + k_{30})x_{10} - k_9x_4x_{10} + k_{31}x_{11} + k_{10}x_{18} \\
\dot{x}_{11} &= -k_{23}x_{11} + k_{30}x_{10} - k_{31}x_{11} - k_{20}x_6x_{11} - k_{24}x_{11}x_{12} + k_{25}x_{13} + k_{21}x_{19} \\
\dot{x}_{12} = -\dot{x}_{13} &= -k_{24}x_{11}x_{12} + k_{25}x_{13} \\
\dot{x}_{14} &= k_3x_2x_4 - (k_4 + k_5)x_{14} \\
\dot{x}_{15} &= k_{14}x_3x_6 - (k_{15} + k_{16})x_{15} \\
\dot{x}_{18} &= k_9x_4x_{10} - (k_{10} + k_{11})x_{18} \\
\dot{x}_{19} &= k_{20}x_6x_{11} - (k_{21} + k_{22})x_{19}
\end{aligned}$$

G–Harrington–Rosen–Sturmfels, Algebraic Systems Biology: A Case Study for the Wnt Pathway, 2016

Biology $\leftrightarrow$ Algebra and Geometry



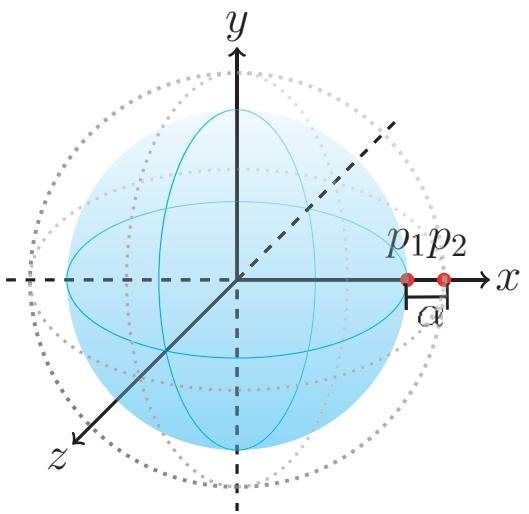
Biology	*Algebra and Geometry*
Multistationarity	Real Algebraic Geometry
Experimental Design	Algebraic Matroids
Model Dynamics	Polyhedral Geometry
Model Selection	Solving Polynomial Systems

Model Selection & Steady State Invariants

A **steady-state invariant** is a polynomial in the species concentrations (the x's) and the rate constants (the k's) that vanishes when the system is at steady state.

Steady-state invariants can be used to perform model selection by

- Comparing the behavior of the species concentrations with the algebraic relation defined by the steady-state invariant (Gunawardena 2007).
- Computing the maximum likelihood using numerical algebraic geometry (G-Davis-Ho-Bates-Harrington 2016)



Open Problems

- **Computing elimination ideals** Elimination ideals are used for model selection. (Exploring how to construct elimination ideals by looking at subnetworks with Heather Harrington, Nikki Meshkat, and Anne Shiu)
- **Steady state degree** The steady-state degree is the number of complex solutions to the steady-state equations for generic choice of parameters. (Ongoing work with Cvetelina Hill).
- **Euclidean distance degree** The ED degree quantifies the algebraic complexity of solving the goodness-of-fit problem. (Current work by Michael Adamer and Martin Helmer)

Thank you!

Thank you!