Identification of cell signaling pathways based on biochemical reaction kinetics repositories

Gustavo Estrela

May 2019

Instituto de Matemática e Estatística
Centro de Toxinas, Resposta-imune e Sinalização Celular (CeTICS)
Laboratório Especial de Ciclo Celular, Instituto Butantan

Introduction

Cell Signaling

Cell signaling allows cells to respond to signals that come from its environment changing its behaviour accordingly.

Cell Signaling

Cell signaling allows cells to respond to signals that come from its environment changing its behaviour accordingly.

This mechanism is essential for many cell functions, including reproduction, growth and death.

Cell Signaling

Cell signaling allows cells to respond to signals that come from its environment changing its behaviour accordingly.

This mechanism is essential for many cell functions, including reproduction, growth and death.

Understanding the functioning of cell signaling is important in many biological areas.

Cell Signaling

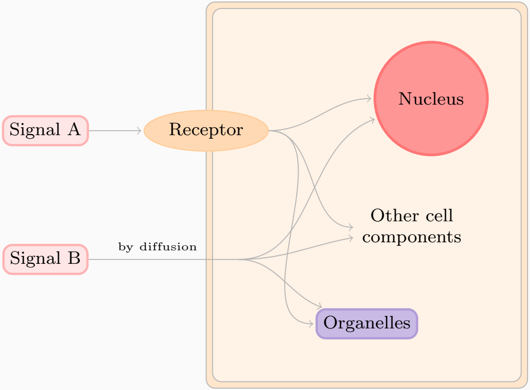


Figura 1: A general cell signaling mechanism.

Cell Signaling Pathways

A cell signaling network can be characterized by a sequence of chemical reactions that allows the presence of a signal to modify the state or behaviour of a cell.

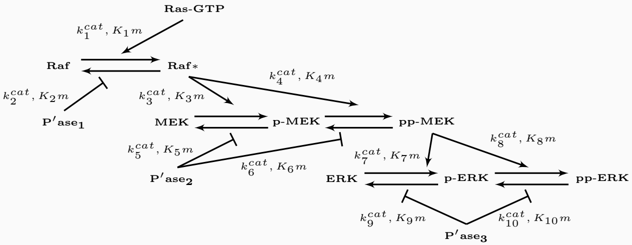


Figura 2: An example of a signaling pathway.

Mathematical Models of Signaling Networks

In many applications, we summarize the state of the cell with a measurement based on the concentration of chemical species of the cell.

Mathematical Models of Signaling Networks

In many applications, we summarize the state of the cell with a measurement based on the concentration of chemical species of the cell.

Using biochemical and enzymatic kinetics, we can write equations that represent the rate of change of concentration for a chemical species.

Mathematical Models of Signaling Networks

As an example, for the chemical reaction

$$A \longrightarrow B,$$

Mathematical Models of Signaling Networks

As an example, for the chemical reaction

$$\mathrm{A} \longrightarrow \mathrm{B},$$

we can write the following equation:

$$\frac{d[\mathsf{A}]}{dt} = k_1[\mathsf{A}];$$

where k_1 is a reaction rate constant.

Mathematical Models of Signaling Networks

As an example, for the chemical reaction

$$\mathrm{A} \longrightarrow \mathrm{B},$$

we can write the following equation:

$$\frac{d[\mathsf{A}]}{dt} = k_1[\mathsf{A}];$$

where k_1 is a reaction rate constant.

Repeating this procedure for all reactions of a pathway allows us to derive a system of ordinary differential equations that can model the signaling pathway.

Identification of Cell Signaling Pathways

The problem of identification of cell signaling pathways is the problem of finding the components of a signaling pathway and how they interact given a set of experimental measurement.

Identification of Cell Signaling Pathways

The problem of identification of cell signaling pathways is the problem of finding the components of a signaling pathway and how they interact given a set of experimental measurement.

As the input, a description of a biological experiment and a set of experimental measurements are given.

Identification of Cell Signaling Pathways

The problem of identification of cell signaling pathways is the problem of finding the components of a signaling pathway and how they interact given a set of experimental measurement.

As the input, a description of a biological experiment and a set of experimental measurements are given. A possible output to the problem is composed by:

Identification of Cell Signaling Pathways

The problem of identification of cell signaling pathways is the problem of finding the components of a signaling pathway and how they interact given a set of experimental measurement.

As the input, a description of a biological experiment and a set of experimental measurements are given. A possible output to the problem is composed by:

- a model composed by a set of chemical reactions that are relevant for the biological experiment;

Identification of Cell Signaling Pathways

The problem of identification of cell signaling pathways is the problem of finding the components of a signaling pathway and how they interact given a set of experimental measurement.

As the input, a description of a biological experiment and a set of experimental measurements are given. A possible output to the problem is composed by:

- a model composed by a set of chemical reactions that are relevant for the biological experiment;
- information about the reaction rate constants of the model.

Identification of Cell Signaling Pathways

One can search for the set of chemical reactions relevant for a biological experiment in repositories like the Kyoto Encyclopedia of Genes and Genomes (KEGG).

Identification of Cell Signaling Pathways

One can search for the set of chemical reactions relevant for a biological experiment in repositories like the Kyoto Encyclopedia of Genes and Genomes (KEGG). However, the pathway maps from KEGG may be incomplete or have impertinent reactions for the biological experiment of interest.

Identification of Cell Signaling Pathways

One can search for the set of chemical reactions relevant for a biological experiment in repositories like the Kyoto Encyclopedia of Genes and Genomes (KEGG). However, the pathway maps from KEGG may be incomplete or have impertinent reactions for the biological experiment of interest.

Hence, it is desirable to construct a method that can systematically modify these models and choose the one that better represents the experiment.

Identification of Cell Signaling Pathways

Lulu Wu (2015) presented in her master dissertation a methodology that proposes to systematically modify models of signaling network in order to better represent experiments.

Identification of Cell Signaling Pathways

Lulu Wu (2015) presented in her master dissertation a methodology that proposes to systematically modify models of signaling network in order to better represent experiments.

On her work, the problem of identification of cell signaling pathways is treated as a feature selection problem.

Feature Selection Problem

The feature selection problem is a combinatorial optimization problem:

> Given a set of features S and a cost function c, find subset $X \in \mathcal{P}(S)$, with minimum cost $c(X)$.

Feature Selection Problem

The feature selection problem is a combinatorial optimization problem:

Given a set of features S and a cost function c, find subset $X \in \mathcal{P}(S)$, with minimum cost $c(X)$.

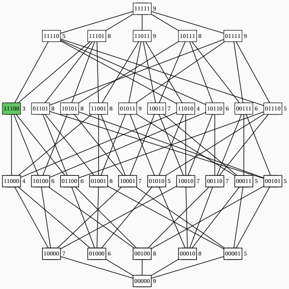


Figura 3: An example of feature selection search space with 5 features.

Feature Selection for Identification of Signaling Pathways

The methodology proposed by Wu defines the set of features as a set of chemical reactions that can be added to a starting model.

Feature Selection for Identification of Signaling Pathways

The methodology proposed by Wu defines the set of features as a set of chemical reactions that can be added to a starting model.This set of chemical reactions is fetched from KEGG and stored in a database of interactions.

Wu's Search Algorithm for Feature Selection

The search algorithm used by Wu is the Sequential Forward Selection (SFS).

Wu's Cost Function for Feature Selection

Wu defines the cost function as the minimum distance between experimental and simulated data.

Wu's Cost Function for Feature Selection

Wu defines the cost function as the minimum distance between experimental and simulated data. The minimum distance is found using a Simmulated Annealing that traverses the parameter space.

Wu's Cost Function for Feature Selection

If we define that:

Wu's Cost Function for Feature Selection

If we define that:

- the experimental measurement is D;

Wu's Cost Function for Feature Selection

If we define that:

- the experimental measurement is D;
- the simulated measure is $\phi(M, \theta)$;

Wu's Cost Function for Feature Selection

If we define that:

- the experimental measurement is D;
- the simulated measure is $\phi(M, \theta)$;
- all elements visited on the Simmulated Annealing process define Θ'.

Wu's Cost Function for Feature Selection

If we define that:

- the experimental measurement is D;
- the simulated measure is $\phi(M, \theta)$;
- all elements visited on the Simmulated Annealing process define Θ'.

Then, the cost function used on Wu's work is:

$$c(M) = min_{\{\theta \in \Theta' \subset \Theta\}} dist(\phi(M, \theta), D) + R(M),$$

Wu's Cost Function for Feature Selection

If we define that:

- the experimental measurement is D;
- the simulated measure is $\phi(M, \theta)$;
- all elements visited on the Simmulated Annealing process define Θ'.

Then, the cost function used on Wu's work is:

$$c(M) = min_{\{\theta \in \Theta' \subset \Theta\}} dist(\phi(M, \theta), D) + R(M),$$

where $R(M)$ is a regularization function.

Wu's Cost Function for Feature Selection

The $R(M)$ term is implicitly defined by imposing a time limit to the Simmulated Annealing procedure used to calculate the cost function.

Wu's Cost Function for Feature Selection

The $R(M)$ term is implicitly defined by imposing a time limit to the Simmulated Annealing procedure used to calculate the cost function. As a result, the penalization of the cost function is random.

Results of Wu's Methodology

Lulu Wu tested her methodology by trying to recreate models given a cut of the original model.

Results of Wu's Methodology

Lulu Wu tested her methodology by trying to recreate models given a cut of the original model. However, the methodology worked satisfactorily only when the cut was similar to the original model.

Difficulties of Wu's Methodology

We can point three aspects of Wu's work that could explain its limitation.

Difficulties of Wu's Methodology

We can point three aspects of Wu's work that could explain its limitation.

- the database of interactions used could be more nearly complete;

Difficulties of Wu's Methodology

We can point three aspects of Wu's work that could explain its limitation.

- the database of interactions used could be more nearly complete;
- the search algorithm could also consider removing interactions;

Difficulties of Wu's Methodology

We can point three aspects of Wu's work that could explain its limitation.

- the database of interactions used could be more nearly complete;
- the search algorithm could also consider removing interactions;
- the cost function could implement a proper penalization of models;

What we Propose on this Project

We propose to create a methodology that uses a feature selection approach for identification of signaling pathways, tackling the difficulties encountered by Wu.

What we Propose on this Project

To get a more nearly complete database of interactions, we should fetch information from KEGG and other databases,

What we Propose on this Project

To get a more nearly complete database of interactions, we should fetch information from KEGG and other databases, such as STRING and SABIO-RK.

What we Propose on this Project

To create new search algorithms,

What we Propose on this Project

To create new search algorithms, we intend to use more general algorithms that can also remove interactions.

What we Propose on this Project

To define new cost functions,

What we Propose on this Project

To define new cost functions, we intend to use Bayesian approaches of model selection that allow us to create estimates of probabilities such as $p(M|D)$ or $p(D|M)$.

Objectives of this Project

Objectives of this Project

- Build a database of interactions.

Objectives of this Project

- Build a database of interactions.
- Create a cost function for models of signaling pathways.

Objectives of this Project

- Build a database of interactions.
- Create a cost function for models of signaling pathways.
- Formulate systematic modifications to a model as the search space of a feature selection model.

Objectives of this Project

- Build a database of interactions.
- Create a cost function for models of signaling pathways.
- Formulate systematic modifications to a model as the search space of a feature selection model.
- Create search algorithms for the feature selection problem.

Objectives of this Project

- Build a database of interactions.
- Create a cost function for models of signaling pathways.
- Formulate systematic modifications to a model as the search space of a feature selection model.
- Create search algorithms for the feature selection problem.
- Test the methodology on known signaling pathways.

Objectives of this Project

- Build a database of interactions.
- Create a cost function for models of signaling pathways.
- Formulate systematic modifications to a model as the search space of a feature selection model.
- Create search algorithms for the feature selection problem.
- Test the methodology on known signaling pathways.
- Apply the methodology on a real case.

Fundamental Concepts

Kinetics Modeling of Chemical Reactions

Mathematical Modeling of Reactions

In this project we use three possible models of kinetics of an interaction:

Mathematical Modeling of Reactions

In this project we use three possible models of kinetics of an interaction:

- first order interaction kinetics;

Mathematical Modeling of Reactions

In this project we use three possible models of kinetics of an interaction:

- first order interaction kinetics;
- second order interaction kinetics;

Mathematical Modeling of Reactions

In this project we use three possible models of kinetics of an interaction:

- first order interaction kinetics;
- second order interaction kinetics;
- Michaelis-Menten enzymatic kinetics.

Kinetic Modeling of First Order Iteration

A first order reaction:

$$\mathbf{A} \xrightarrow{k_1} \mathbf{B}$$

Kinetic Modeling of First Order Iteration

A first order reaction:

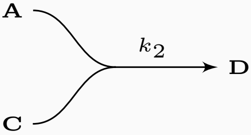
$$\mathbf{A} \xrightarrow{k_1} \mathbf{B}$$

has rate of:

$$k_1[\mathrm{A}].$$

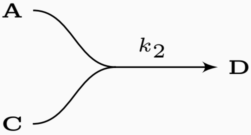
Kinetic Modeling of Second Order Iteration

A second order reaction:



Kinetic Modeling of Second Order Iteration

A second order reaction:

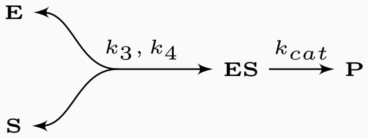


has rate of:

$$k_1[\mathrm{A}][\mathrm{C}].$$

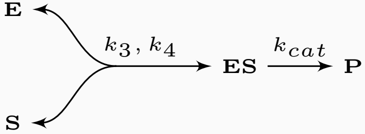
Kinetic Modeling of Enzymatic Reactions

An enzymatic reaction:



Kinetic Modeling of Enzymatic Reactions

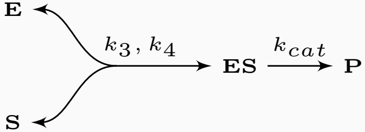
An enzymatic reaction:



Can be divided in two first order reactions plus a second order reaction.

Kinetic Modeling of Enzymatic Reactions

An enzymatic reaction:



Can be divided in two first order reactions plus a second order reaction. However, with the appropriate assumptions, it is possible to use a Michaelis-Menten simplification of this reaction.

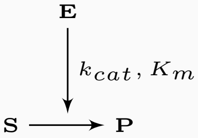
Michaelis-Menten Kinetics

We denote Michaelis-Menten simplification of the last enzymatic reaction as

$$\begin{array}{c} \mathbf{E} \\ \Big\downarrow {\scriptstyle k_{cat}\,,\,K_m} \\ \mathbf{S} \longrightarrow \mathbf{P} \end{array}$$

Michaelis-Menten Kinetics

We denote Michaelis-Menten simplification of the last enzymatic reaction as

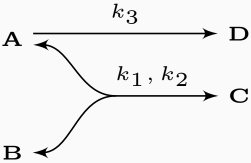


and it has rate of:

$$k_{cat}\frac{[\mathrm{E}][\mathrm{S}]}{K_M + [\mathrm{S}]}.$$

Kinetics of a System of Reactions

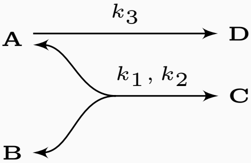
Suppose we want to model the kinetics of A on these reactions:



This system can be divided in three reactions:

Kinetics of a System of Reactions

Suppose we want to model the kinetics of A on these reactions:

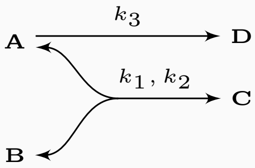


This system can be divided in three reactions:

- A + B $\longrightarrow$ C, with rate k_1[A][B],

Kinetics of a System of Reactions

Suppose we want to model the kinetics of A on these reactions:

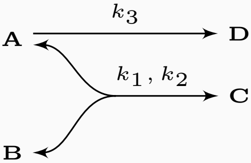


This system can be divided in three reactions:

- $A + B \longrightarrow C$, with rate $k_1[A][B]$,
- $C \longrightarrow A + B$, with rate $k_2[C]$,

Kinetics of a System of Reactions

Suppose we want to model the kinetics of A on these reactions:



This system can be divided in three reactions:

- $A + B \longrightarrow C$, with rate $k_1[A][B]$,
- $C \longrightarrow A + B$, with rate $k_2[C]$,
- $A \longrightarrow D$, with rate $k_3[A]$.

Kinetics of a System of Reactions

In $A + B \longrightarrow C$, with rate $k_1[A][B]$, A is a reactant.

Kinetics of a System of Reactions

In $\mathrm{A} + \mathrm{B} \longrightarrow \mathrm{C}$, with rate $k_1[\mathrm{A}][\mathrm{B}]$, A is a reactant.

In $\mathrm{C} \longrightarrow \mathrm{A} + \mathrm{B}$, with rate $k_2[\mathrm{C}]$, A is a product.

Kinetics of a System of Reactions

In $A + B \longrightarrow C$, with rate $k_1[A][B]$, A is a reactant.

In $C \longrightarrow A + B$, with rate $k_2[C]$, A is a product.

In $A \longrightarrow D$, with rate $k_3[A]$, A is a reactant.

Kinetics of a System of Reactions

In $\mathrm{A} + \mathrm{B} \longrightarrow \mathrm{C}$, with rate $k_1[\mathrm{A}][\mathrm{B}]$, A is a reactant.

In $\mathrm{C} \longrightarrow \mathrm{A} + \mathrm{B}$, with rate $k_2[\mathrm{C}]$, A is a product.

In $\mathrm{A} \longrightarrow \mathrm{D}$, with rate $k_3[\mathrm{A}]$, A is a reactant.

Then, the differential equation that models the concentration change of A is:

Kinetics of a System of Reactions

In $\mathrm{A} + \mathrm{B} \longrightarrow \mathrm{C}$, with rate $k_1[\mathrm{A}][\mathrm{B}]$, A is a reactant.

In $\mathrm{C} \longrightarrow \mathrm{A} + \mathrm{B}$, with rate $k_2[\mathrm{C}]$, A is a product.

In $\mathrm{A} \longrightarrow \mathrm{D}$, with rate $k_3[\mathrm{A}]$, A is a reactant.

Then, the differential equation that models the concentration change of A is:

$$\frac{d[\mathrm{A}]}{dt} = -k_1[\mathrm{A}][\mathrm{B}] + k_2[\mathrm{C}] - k_3[\mathrm{D}].$$

Bayesian Methods for Biochemical Model Selection

State of the Art Methods for Model Selection

There are two main Bayesian methods available for biochemical model selection:

State of the Art Methods for Model Selection

There are two main Bayesian methods available for biochemical model selection:

- Approximate Bayesian Computation;

State of the Art Methods for Model Selection

There are two main Bayesian methods available for biochemical model selection:

- Approximate Bayesian Computation;
- Marginal likelihood estimation through Thermodynamic Integration.

State of the Art Methods for Model Selection

There are two main Bayesian methods available for biochemical model selection:

- Approximate Bayesian Computation;
- Marginal likelihood estimation through Thermodynamic Integration.

For both methods, we resort to Metropolis-Hastings algorithm to generate samples of distributions.

Metropolis-Hastings algorithm

With Metropolis-Hastings, we can generate a sample of a distribution $p(\lambda)$ doing the following:

Metropolis-Hastings algorithm

With Metropolis-Hastings, we can generate a sample of a distribution $p(\lambda)$ doing the following:

Metropolis-Hastings algorithm

With Metropolis-Hastings, we can generate a sample of a distribution $p(\lambda)$ doing the following:

1. Choose some λ_0 for which $p(\lambda_0) > 0$, and set $t = 1$;

Metropolis-Hastings algorithm

With Metropolis-Hastings, we can generate a sample of a distribution $p(\lambda)$ doing the following:

1. Choose some λ_0 for which $p(\lambda_0) > 0$, and set $t = 1$;
2. Sample a candidate point λ^* from a jump distribution, $J(\lambda|\lambda_{t-1})$;

Metropolis-Hastings algorithm

With Metropolis-Hastings, we can generate a sample of a distribution $p(\lambda)$ doing the following:

1. Choose some λ_0 for which $p(\lambda_0) > 0$, and set $t = 1$;
2. Sample a candidate point λ^* from a jump distribution, $J(\lambda|\lambda_{t-1})$;
3. Calculate the ratio $r = \frac{p(\lambda^*)J_t(\lambda^{t-1}|\lambda^*)}{p(\lambda^{t-1})J_t(\lambda^*|\lambda^{t-1})}$;

Metropolis-Hastings algorithm

With Metropolis-Hastings, we can generate a sample of a distribution $p(\lambda)$ doing the following:

1. Choose some λ_0 for which $p(\lambda_0) > 0$, and set $t = 1$;
2. Sample a candidate point λ^* from a jump distribution, $J(\lambda|\lambda_{t-1})$;
3. Calculate the ratio $r = \frac{p(\lambda^*)J_t(\lambda^{t-1}|\lambda^*)}{p(\lambda^{t-1})J_t(\lambda^*|\lambda^{t-1})}$;
4. Set $\lambda_t = \lambda^*$ with probability $\min(1, r)$ and $\lambda_t = \lambda_{t-1}$ otherwise;

Metropolis-Hastings algorithm

With Metropolis-Hastings, we can generate a sample of a distribution $p(\lambda)$ doing the following:

1. Choose some λ_0 for which $p(\lambda_0) > 0$, and set $t = 1$;
2. Sample a candidate point λ^* from a jump distribution, $J(\lambda|\lambda_{t-1})$;
3. Calculate the ratio $r = \frac{p(\lambda^*)J_t(\lambda^{t-1}|\lambda^*)}{p(\lambda^{t-1})J_t(\lambda^*|\lambda^{t-1})}$;
4. Set $\lambda_t = \lambda^*$ with probability $\min(1, r)$ and $\lambda_t = \lambda_{t-1}$ otherwise;
5. Increase t by one and repeat from Step 2 if not reached iteration limit.

Model Selection

Experiments on Model Selection

Next Steps