# Identificação de vias de sinalização celular baseada em repositórios de cinética de reações bioquímicas

**Bolsista:** Gustavo Estrela de Matos

**Orientador:** Marcelo da Silva Reis

Centro de Toxinas, Imuno-resposta e Sinalização Celular (CeTICS)

Laboratório Especial de Ciclo Celular (LECC)

Instituto Butantan, São Paulo, 25 de setembro de 2017.

**Resumo**

[A fazer (máximo 20 linhas).]

# Identification of cell signaling pathways based on biochemical reaction kinetics repositories

**Student:** Gustavo Estrela de Matos

**Supervisor:** Marcelo da Silva Reis

Center of Toxins, Immune-response and Cell Signaling (CeTICS)

Laboratório Especial de Toxinologia Aplicada (LETA)

Instituto Butantan, São Paulo, September 25, 2017.

**Abstract**

[To do (20 lines maximum).]

# Sumário

# 1 Introdução

A sinalização celular é o processo de troca de informações que permite às células interagir com o ambiente por meio de interações entre espécies químicas. Chamamos um conjunto de interações desse tipo que estão associadas a uma função celular de uma via de sinalização celular. A sinalização celular coordena diversos processos da célula e entender a topologia de suas vias pode ajudar a melhorar o entendimento de processos celulares assim como permitir criar novos tratamentos para doenças como o câncer e diabetes, que podem ser causadas por anomalias nestas vias.

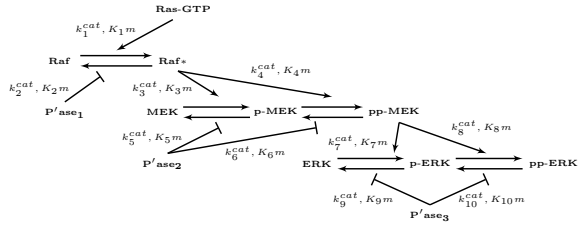
Estudamos uma via de sinalização observando a concentração de espécies químicas ao longo do tempo, o que pode ser feito através de experimentos biológicos ou com modelos computacionais. Estes modelos, também chamados de modelos funcionais, podem descrever a concentração de espécies químicas ao longo do tempo por sistemas de equações diferenciais, seguindo as regras definidas pela cinética de reações bioquímicas. Os modelos funcionais são apenas capazes de simular o que pode ser observado por experimentos biológicos e a vantagem dessa abordagem consiste em, considerando que o modelo aproxima bem a realidade, prever o estado da célula dado um estado inicial e uma porção de tempo.

## 1.1 Identificação de vias de sinalização

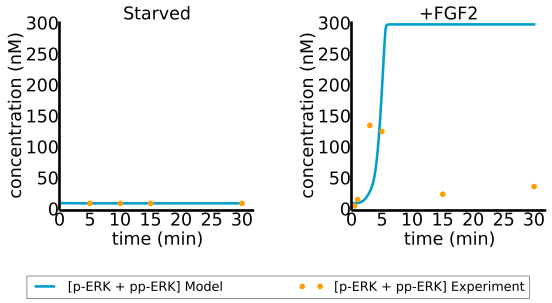
O problema de desenhar modelos funcionais é chamado de identificação de vias de sinalização celular e esse problema apresenta duas dificuldades principais: a primeira é escolher as espécies químicas e interações que participam da via, e a segunda é determinar constantes que aparecem no sistema de equações diferenciais das concentrações de espécies químicas. Para a primeira dificuldade, podemos usar como base recortes de mapas de interatomas disponíveis em bancos de dados como o Kyoto Encyclopedia of Genes and Genomes (KEGG) [3]. Já para a segunda dificuldade, podemos usar constantes disponíveis na literatura ou determinar valores fazendo o ajuste da curva (do inglês *curve fitting*) de dados gerados pelo

modelo funcional pelos dados coletados experimentalmente; uma opção de software capaz de fazer tal ajuste de curva é o SigNetSim [17].

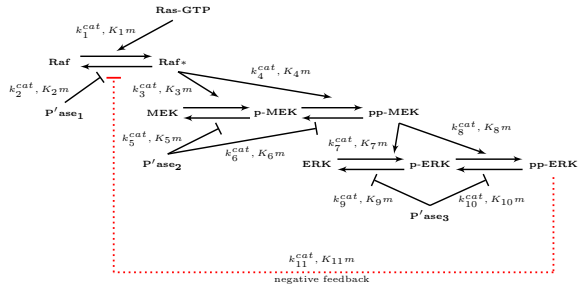
Determinar as espécies químicas e interações de um modelo a partir de um mapa de interatomas torna-se problemático quando o mapa é muito grande ou quando é incompleto. A figura 1 mostra um caso de especificação de modelo funcional em que a escolha de um recorte de um mapa de interatoma do banco de dados KEGG não foi suficiente para que o modelo descrevesse corretamente os experimentos biológicos, o que foi contornado por Reis et al. ao adicionar uma nova interação, que não pertencia ao mapa usado como base [1]. Desta forma, torna-se necessário sistematizar a escolha de espécies químicas e interações de um modelo funcional.



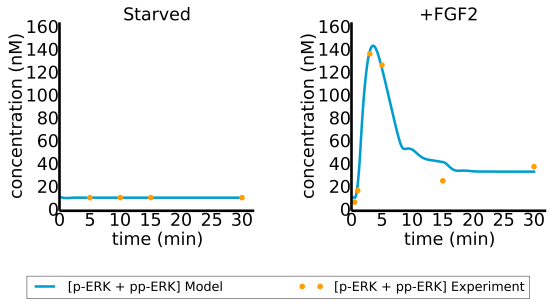

(a)




(b)




(c)




(d)

Figura 1: Exemplo de identificação da via de sinalização Ras/ERK em células carcinoma adrenocorticais Y1 de rato [1]. As figuras 1(a) e 1(c) representam respectivamente o modelo funcional e um gráfico comparando dados da simulação e do experimento biológico. O gráfico mostra que a simulação não se aproxima bem da realidade, o que é contornado por Reis et al. adicionando uma nova interação no modelo funcional (em vermelho na figura 1(b)) . O modelo atualizado e os resultados da simulação são apresentados respectivamente nas figuras 1(b) e 1(d)

## 1.2 Modificação de modelos funcionais a partir de bancos de dados de interatomas

Com intuito de sistematizar a modificação de um modelo funcional, foi desenvolvido em uma tese de mestrado uma abordagem de identificação de vias de sinalização celular que aumenta recortes de mapas de interatomas com o objetivo de que o modelo represente corretamente os dados biológicos [2]. Esta abordagem modifica os modelos funcionais de forma incremental, adicionando interações de um banco de dados, que foi construído com a união de várias interações de diferentes mapas de interatomas do KEGG.

Mais formalmente, esta abordagem trata o problema de identificação de vias de sinalização como um problema de otimização, em que o espaço de busca é constituído por todos os modelos que podem ser construídos a partir do modelo original (incluindo ele mesmo), aumentado com interações do banco de dados usado. A função de custo é dada pelo erro do modelo no ajuste de curva dos dados biológicos, calculado pelo software SigNetSim; a busca pelo melhor elemento do espaço de busca é feita usando o algoritmo de busca *Sequential Forward Selection* (SFS).

Esta abordagem, entretanto, apresentou algumas limitações. A primeira é a incompletude do banco de interações criado, que usava apenas mapas de interatomas disponíveis no KEGG e, além disso, não considera constantes de velocidade disponíveis em outros bancos de dados, como o Sabio-RK [12], o que aumenta desnecessariamente a complexidade do modelo funcional. A segunda limitação está no algoritmo de busca, que por ser incremental pode "caminhar" o modelo até um mínimo local, perdendo a melhor solução. A terceira e última está na penalização de modelos mais complexos, que era feita implicitamente ao impor um tempo de limite no ajuste de curva do modelo, ignorando a velocidade de convergência do algoritmo para o modelo em questão, resultando em uma penalização aleatória.

# 2 Objetivos

- Geral: desenvolver uma abordagem mais efetiva para auxiliar na identificação de vias de sinalização celular. Esta abordagem teria como ponto de partida o trabalho da Lulu e incluiria soluções para as limitações do mesmo, que foram discutidas na seção anterior.

- Específico: aplicar a metodologia na identificação de vias de sinalização celular relevantes em nosso estudo de caso, a linhagem tumoral murina Y1.

# 3 Metodologia

## 3.1 Desafios científicos

1. Realizar a seleção de modelos utilizando uma estratégia global ao invés de incremental. Para este fim, faremos a redução do problema da seleção de modelos para um problema de seleção de características, o que exigirá:

    - Definir uma função custo apropriada, que leve em consideração a penalização por *overfitting* decorrente do acréscimo de novas espécies químicas e/ou reações sem a inclusão de novas medidas experimentais para ajustar o modelo aumentado. Uma possibilidade seria o uso do critério de informação de Akaike (*Akaike's Information Criterion* – AIC) [5], cujo princípio foi aplicado com sucesso em seleção de modelos no contexto de discriminação de classes de redes biológicas [6]. Também investigaremos para este fim o uso de abordagens Bayesianas [7], como por exemplo a técnica conhecida como *Bayesian inference-based modeling* (BIBm) [8].

    - Escolher um algoritmo de seleção de características. Critérios que penalizam *overfitting* provavelmente induzirão curvas em U nas cadeias do reticulado Booleano induzido pelo espaço de busca, o que nos permitirá aproximar o problema de seleção de características através do problema de otimização U-curve. Como o

cálculo da função custo provavelmente será computacionalmente muito intensivo, o melhor algoritmo para esse fim tende a ser o U-Curve-Search (UCS) [9, 10].

2. Contornar o problema da incompletude dos bancos de dados de interatomas (e.g., KEGG), que se dá no nível de estrutura da via de sinalização e também na ausência de constantes de velocidade.

   - **Estrutura da via de sinalização.** no mapa da via de sinalização de Ras em camundongo, existe uma aresta dizendo que Raf1 ativa MEK, mas não como se dá tal ativação. Por exemplo, no modelo apresentado na introdução, MEK é ativado após ser fosforilado duas vezes pela forma ativa de Raf1, dinâmica que exige duas equações para ser descrita em termos de cinética química. Para lidar com este problema, precisaremos estabelecer a lei cinética (tipo de reação) a ser utilizada de acordo com a natureza das espécies químicas envolvidas em uma dada interação. No exemplo dado, sabe-se que a ativação de MEK por Raf é ultrassensível, e que portanto pode ser modelada utilizando a equação de Hill [11].

   - **Ausência de constantes de velocidade.** Dificulta tanto a estimação quanto o problema de otimização. Vamos contornar isso coletando e organizando informações extraídas de repositórios tais como o Sabio-RK [12], Brenda [13], BioNumbers [14], e possivelmente também BioModels [15].

## 3.2 Desafios tecnológicos

1. Integração apropriada do featsel [16] com o SigNetSim [17] para seleção de modelos.

2. Organização das informações coletadas em um banco de dados relacional, que será integrado ao CeTICSdb, repositório de ômicas desenvolvido e mantido pelo grupo de Biologia Computacional do CeTICS.

# 4 Plano de trabalho e cronograma de execução

[Tente detalhar os trabalhos que serão necessários, dados os objetivos e desafios metodológicos.]

## 4.1 Cronograma proposto

[Tabela análoga ao dos projetos anteriores.]

# 5 Forma de análise e disseminação de resultados

[Análogo ao nosso último projeto.]

# Referências


[1] Marcelo S Reis, Vincent Noël, Matheus H Dias, Layra L Albuquerque, Amanda S Guimarães, Lulu Wu, Junior Barrera, and Hugo A Armelin. An Interdisciplinary Approach for Designing Kinetic Models of the Ras/MAPK Signaling Pathway. In *Kinase Signaling Networks*, pages 455–474. Springer, 2017.

[2] Lulu Wu. Um método para modificar vias de sinalização molecular por meio de análise de banco de dados de interatomas. Master's thesis, Universidade de São Paulo, 2015.

[3] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[4] A. W. Whitney. A direct method of nonparametric measurement selection. *IEEE Trans Comp*, 20(9):1100–1103, 1971.

[5] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[6] Daniel Yasumasa Takahashi, Joao Ricardo Sato, Carlos Eduardo Ferreira, and André Fujita. Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PLoS One*, 7(12):e49949, 2012.

[7] Paul Kirk, Thomas Thorne, and Michael PH Stumpf. Model selection in systems and synthetic biology. *Current opinion in biotechnology*, 24(4):767–774, 2013.

[8] Tian-Rui Xu, Vladislav Vyshemirsky, Amélie Gormand, Alex von Kriegsheim, Mark Girolami, George S. Baillie, Dominic Ketley, Allan J. Dunlop, Graeme Milligan, Miles D. Houslay, and Walter Kolch. Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Science Signaling*, 3(113):ra20–ra20, 2010.

[9] Marcelo S. Reis. *Minimização de funções decomponíveis em curvas em U definidas sobre cadeias de posets–algoritmos e aplicações*. PhD thesis, Universidade de São Paulo, 2012.

[10] Marcelo S. Reis, Gustavo Estrela, Carlos Eduardo Ferreira, and Junior Barrera. Optimal Boolean lattice-based algorithms for the U-curve optimization problem. *Enviado para publicação*, 2017.

[11] Chi-Ying Huang and James E Ferrell. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences*, 93(19):10078–10083, 1996.

[12] Ulrike Wittig, Renate Kania, Martin Golebiewski, Maja Rey, Lei Shi, Lenneke Jong, Enkhjargal Algaa, Andreas Weidemann, Heidrun Sauer-Danzwith, Saqib Mir, Olga Krebs, Meik Bittkowski, Elina Wetsch, Isabel Rojas, and Wolfgang Müller. Sabio-rk—database for biochemical reaction kinetics. *Nucleic Acids Research*, 40(D1):D790–D796, 2012.

[13] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32(suppl 1):D431–D433, 2004.

[14] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research*, 38(suppl_1):D750–D753, 2009.

[15] Nicolas Le Novere, Benjamin Bornstein, Alexander Broicher, Melanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert Sauro, Maria Schilstra, Bruce Shapiro, et al. Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic acids research*, 34(suppl_1):D689–D691, 2006.

[16] Marcelo S. Reis, Gustavo Estrela, Carlos Eduardo Ferreira, and Junior Barrera. featsel: A framework for benchmarking of feature selection algorithms and cost functions. *SoftwareX*, 6:193 – 197, 2017.

[17] Vincent Noël. The Signaling Network Simulator (SigNetSim): A Web application for building, fitting, and analyzing mathematical models of molecular signaling networks., 2017. Accessed September 20, 2017. `https://github.com/vincent-noel/SigNetSim`.