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, 26 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 Ciclo Celular (LECC)

Instituto Butantan, São Paulo, September 26, 2017.

Abstract

[To do (20 lines maximum).]

Sumário

1 Introdução

A sinalização celular é um processo de troca de informações que ocorre no interior de uma célula e em suas imediações por meio de interações entre espécies químicas. Um conjunto de interações que estão associadas a uma determinada transmissão de informação (e.g., um sinal que chega no núcleo celular a partir do acionamento de um receptor de membrana citoplasmática) é chamado de via de sinalização celular. Diversos processos celulares são coordenados por vias de sinalização celular; por exemplo, o ciclo celular pode ser estimulado a partir de uma sinalização mitogênica, que trafega por algumas dessas vias (e.g., pelas vias das MAP quinases). Portanto, entender a topologia e a dinâmica dessas vias pode ajudar a melhorar o entendimento do funcionamento de diversos tipos de células. Ademais, uma vez que anomalias em vias de sinalização celular podem levar ao desenvolvimento de doenças tais como o câncer e o diabetes, desvendar propriedades de seus mecanismos é uma etapa inicial, porém altamente relevante, no desenvolvimento de novos tratamentos contra essas doenças.

O estudo de uma via de sinalização é realizado através da observação, ao longo de uma determinada janela de tempo, da concentração de algumas das espécies químicas envolvidas no fenômeno. Dependendo do contexto, a via de sinalização pode ser estudada através da aplicação de testes estatísticos e de análises de correlações sobre resultados dos experimentos biológicos. Todavia, como geralmente é possível medir apenas alguns instantes de tempo de uma fração das espécies químicas envolvidas, faz-se necessário agregar nessas análises informações *a priori* vindas do conhecimento existente da cinética de reações bioquímicas, o que pode ser feito através de modelos dinâmicos computacionais.

Um tipo particular de modelo dinâmico computacional, denominado modelo funcional, pode descrever a concentração de espécies químicas ao longo do tempo, através de algum formalismo matemático das regras definidas pela cinética de reações bioquímicas; por exemplo,

empregando sistemas de equações diferenciais ordinárias (EDOs). Os modelos funcionais, quando corretamente definidos, são capazes de simular o que pode ser observado por experimentos biológicos. Além de incorporar às análises informações *a priori* relevantes, o uso desse tipo de modelo também traz a vantagem de ser uma abordagem preditiva: considerando que o modelo desenhado aproxima bem a realidade e que não padece de *overfitting*, o mesmo pode prever a dinâmica da via de sinalização para diferentes estados (condições) iniciais. Isso pode ser utilizado, por exemplo, para testar o comportamento de uma via de sinalização celular de acordo com o tratamento dado às células.

Portanto, o desenho de modelos funcionais para o estudo de vias de sinalização celular é um problema relevante no contexto de Biologia Celular Molecular e de Biomedicina; uma definição desse problema, bem como métodos para abordá-lo, serão apresentados a seguir.

1.1 Identificação de vias de sinalização celular

O problema de desenhar modelos funcionais, que sejam capazes de explicar os resultados obtidos em experimentos biológicos e que ao mesmo tempo minimizem o problema de *overfitting*, é chamado de *problema de identificação de vias de sinalização celular*. Tipicamente, a resolução desse problema é realizada em duas etapas: na primeira delas, escolhemos as espécies químicas e interações que participam da via, definindo assim as EDOs que farão a descrição matemática da dinâmica do modelo ao longo de uma determinada janela de tempo. Já na segunda etapa, são determinados valores para as constantes de velocidade (i.e., os valores dos parâmetros do sistema de EDOs) e também para as concentrações iniciais (i.e., o estado inicial do modelo dinâmico); se não é possível encontrar valores adequados para os parâmetros, então é preciso voltar para a primeira etapa e redefinir as espécies químicas e/ou interações envolvidas.

Para a resolução da primeira etapa, frequentemente recorre-se a interações entre espécies químicas listadas em bancos de dados de interatomas tais como o Kyoto Encyclopedia of

Genes and Genomes (KEGG) [1]. KEGG e outros bancos de dados similares apresentam uma vasta coleção de interatomas (também conhecidos como mapas estáticos), catalogados por organismo, tipo de tecido, via de sinalização estudada, etc. Logo, o pesquisador normalmente recorre ao mapa cuja categoria mais se aproxime das condições de seu experimento biológico para selecionar as interações que comporão seu modelo funcional. Outra alternativa é recorrer a um modelo *scaffolding* para servir de base ao modelo funcional em desenvolvimento, para este fim recorrendo a bancos de dados de modelos; um exemplo desse tipo de banco de dados é o BioModels [2].

Já para a segunda etapa, é possível utilizar constantes de velocidade disponíveis na literatura e/ou em bancos de dados tal como o BioNumbers [3], que são determinadas através de experimentos biológicos (e.g., em ensaios enzimáticos). Quando tais constantes não estão disponíveis, precisamos recorrer a um método conhecido como otimização por ajuste de curva (do inglês *curve-fitting optimization*). Nesse tipo de otimização, para um dado conjunto de valores para os parâmetros do modelo funcional, uma simulação é realizada e seu resultado é comparado com os dados coletados experimentalmente, segundo uma dada medida de erro (i.e., a função custo); com a ajuda de um algoritmo de otimização, tal procedimento é realizado muitas vezes, até que algum critério de parada seja atingido (e.g., um limiar superior para o erro). Uma opção de ferramenta para a realização de otimização por ajuste de curva é o Signaling Network Simulator (SigNetSim) [4].

Um exemplo de resolução do problema de identificação de vias de sinalização celular é mostrado na figura 1, no qual é realizada uma primeira iteração da primeira e segunda etapas (figuras 1(a) e 1(b)) que não foi capaz de explicar as medidas dos experimentos biológicos; dessa forma, foi realizada uma segunda iteração dessas duas etapas (figuras 1(c) e 1(d)), com resultados muito mais satisfatórios.

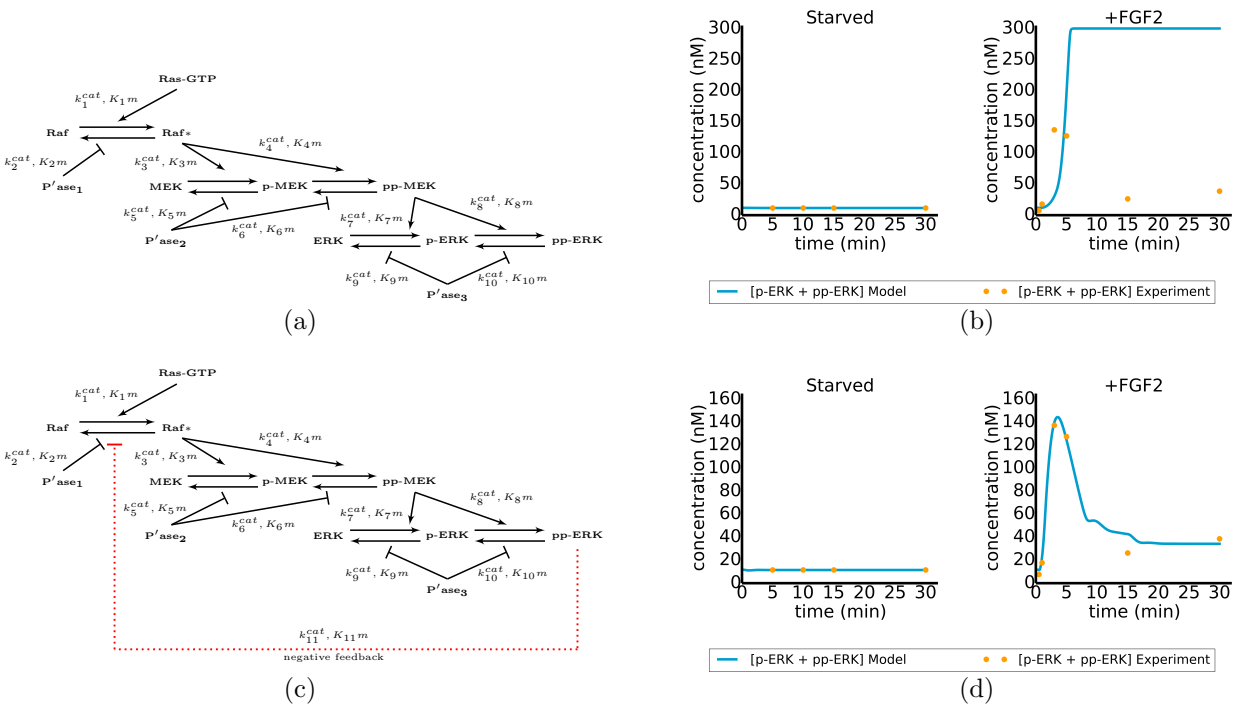

Figura 1: Exemplo de identificação da via de sinalização Ras/ERK em células murinas de carcinoma adrenocortical Y1, conforme apresentado em Reis e colegas [5]. As figuras 1(a) e 1(b) representam, respectivamente, uma hipótese inicial para o modelo funcional e dois gráficos comparando dados da simulação desse modelo com as medidas obtidas no experimento biológico, mostrando que a simulação com esse modelo não se aproxima bem dos valores medidos. Isso foi contornado adicionando uma nova reação bioquímica no modelo funcional (em vermelho na figura 1(c)); a simulação do modelo atualizado resultou em uma cinética que explica os resultados experimentais (figura 1(d)).

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

No exemplo de identificação de vias de sinalização celular mostrado na figura 1, como a via em questão tratava-se da clássica Ras/ERK, não foi difícil localizar na literatura uma reação bioquímica suficiente para completar o modelo funcional. Todavia, em casos em que a via de sinalização celular é pouco ou nada estudada, não é possível recorrer à literatura para buscar espécies e/ou reações químicas para completar o modelo. Neste caso, uma alternativa seria recorrer a mapas estáticos contidos em bancos de dados de interatomas para buscar

espécies e/ou reações químicas candidatas a completar o modelo funcional. Todavia, esta estratégia pode tornar-se problemática quando o mapa é muito grande e/ou quando o mesmo é incompleto.

Com intuito de sistematizar a modificação de um modelo funcional, Lulu Wu introduziu em 2015, em sua dissertação de mestrado pelo IME-USP, uma abordagem de identificação de vias de sinalização celular com o auxílio do banco de dados de interatomas KEGG [6]. Nessa abordagem, todos os mapas estáticos presentes no KEGG que eram referentes a vias de sinalização celular de um dado organismo foram coletados e organizados em um grafo, resultado da união de todos esses mapas. Além disso, durante a primeira etapa da resolução do problema de identificação de vias de sinalização celular, a modificação de um modelo funcional era realizada de forma incremental, adicionando interações presentes nesse grafo; posteriormente, a segunda etapa era realizada utilizando a ferramenta SigNetSim para otimização de ajuste de curva. Em outras palavras, podemos dizer que tal abordagem trata o problema como um problema de otimização, no qual 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 presentes no grafo. A função de custo é dada pelo erro do modelo no ajuste da curva simulada comparada com os dados biológicos, e é calculado pela ferramenta SigNetSim; a busca pelo melhor elemento do espaço de busca é feita usando o algoritmo de busca *Sequential Forward Selection* (SFS) [7].

Esta abordagem, entretanto, apresentou algumas limitações. A primeira delas é a incompletude do banco de interações criado, que usava apenas mapas de interatomas disponíveis no KEGG, não considerando informações complementares que poderiam ser extraídas de bancos de dados bem mais abrangentes para este fim, como por exemplo o STRING [8]. Além disso, a metodologia não considerou as constantes de velocidade que encontram-se disponíveis em bancos de dados de cinética de reações bioquímicas (e.g., o Sabio-RK [9]); não utilizar constantes de velocidades que já tenham sido determinadas experimentalmente au-

menta de forma considerável a dificuldade da otimização por ajuste de curva, além de elevar o risco de *overfitting* do modelo funcional produzido. A terceira limitação está no algoritmo de busca, que por ser incremental pode "caminhar" o modelo até um mínimo local, perdendo assim a melhor solução. Por fim, uma última limitação está na penalização de *overfitting* para 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) [10], cujo princípio foi aplicado com sucesso em seleção de modelos no contexto de discriminação de classes de redes biológicas [11]. Também investigaremos para este fim o uso de abordagens Bayesianas [12], como por exemplo a técnica conhecida como *Bayesian inference-based modeling* (BIBm) [13].

- 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) [14, 15].

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, assim como tratar a 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 [16].
   - **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 [9], Brenda [17], BioNumbers [3], e possivelmente também BioModels [2].

3.2 Desafios tecnológicos

1. Integração apropriada do featsel [18] com o SigNetSim [4] 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] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[2] 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.

[3] 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.

[4] 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`.

[5] 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.

[6] 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.

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

[8] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl_1):D561–D568, 2010.

[9] 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. Sabiork—database for biochemical reaction kinetics. *Nucleic Acids Research*, 40(D1):D790–D796, 2012.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.