Relatório Científico Parcial – Mestrado

Processo FAPESP 17/20575-9

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

Beneficiário: Gustavo Estrela de Matos

Responsável: Marcelo da Silva Reis

Relatório referente aos trabalhos desenvolvidos entre 1 de janeiro e

10 de dezembro de 2018

Laboratório Especial de Toxinologia Aplicada, Instituto Butantan

São Paulo, 7 de Dezembro de 2018

Conteúdo

1 Resumo do Projeto Proposto

A construção de modelos funcionais é uma técnica comum para se estudar vias de sinalização celular e, quando a via estudada é pouco conhecida, é possível que os modelos já propostos sejam incompletos, tornando necessário a sua modificação. Lulu Wu apresentou em 2015, em sua dissertação de mestrado, um método para modificar sistematicamente modelos funcionais, adicionando a estes interações extraídas de repositórios como KEGG. Entretanto, esta metodologia apresentou limitações: a primeira é a incompletude do banco de dados de interações criado, que extraia informações apenas do repositório KEGG; a segunda, a falta de informações sobre constantes de velocidade de interações, que podem ser extraídas de repositórios como BioNumbers; a terceira, a dinâmica do algoritmo de busca, incremental, que pode não achar o mínimo global; e a última, a penalização na complexidade dos modelos, que era feita de maneira aleatória. Propomos neste trabalho enfrentar as limitações encontradas pela metodologia de Lulu, criando um banco de dados de interações mais completo e também novas funções de custo que sejam capazes de penalizar modelos mais complexos (como critério de informação Akaike e *Bayesian inference-based modeling*); esta penalização deve induzir, em cadeias do espaço de busca, curvas em u no custo dos modelos, portanto também propomos a criação de novos algoritmos de busca que explorem essa característica da função de custo. Por fim, esperamos testar nossa metodologia na identificação de vias de sinalização celular da linhagem tumoral murina Y1.

2 Introdução

Vias de sinalização celular podem ser simuladas por modelos dinâmicos computacionais e, mais especificamente, modelos que descrevem a concentração de espécies químicas ao longo do tempo são chamados de modelos funcionais. Neste projeto, trabalhamos com modelos funcionais que descrevem as mudanças de concentrações de espécies químicas através de equações diferenciais ordinárias (EDOs). Estes modelos, quando não sofrem de sobreajuste

(*overfitting*), se tornam interessantes quando são capazes de reproduzir dados observados em experimentos biológicos, pois dão a estes modelos a qualidade preditiva.

O problema de criar modelos funcionais capazes de explicar resultados de experimentos biológicos com o mínimo de sobreajuste é chamado de *problema de identificação de vias de sinalização celular*. Podemos separar este problema em duas etapas principais. A primeira diz respeito a escolha da topologia da via de sinalização, o que é equivalente a escolher quais interações químicas são relevantes para o experimento em questão. A segunda etapa consiste em escolher valores para os parâmetros do sistema de EDOs do modelo funcional; estes valores são constantes de velocidade de interações e/ou concentrações iniciais de espécies químicas.

Em casos em que o experimento biológico ou a via de sinalização são muito estudadas, é possível que ambas as etapas descritas anteriormente possam ser resolvidas com pesquisas na literatura. Em casos que isto não é possível, torna-se uma solução recorrer a bancos de dados de biologia. Para a primeira etapa, podemos consultar bancos como o Kyoto Encyclopedia of Genes and Genomes (KEGG) [1], que contém mapas estáticos; e para a segunda, podemos consultar bancos como o BioModels [2] .

Entretanto, os mapas estáticos disponíveis nestes bancos podem ainda ser incompletos ou muito grandes para o experimento biológico em questão. Desta maneira, torna-se importante criar uma maneira sistemática de se modificar modelos funcionais a fim de faze-los explicar o experimento biológico sem sobreajuste. Propomos fazer estas modificações tratando o problema de identificação de vias de sinalização como um problema de otimização combinatória, considerando como espaço de busca possíveis topologias para o modelo funcional e usando como função de custo alguma métrica que represente a qualidade deste modelo ao reproduzir o experimento biológico de interesse.

Considere S um conjunto de interações químicas. O conjunto de todas as possíveis escolhas de interações relevantes em S corresponde ao conjunto potência de S, $\mathcal{P}(S)$, que é o espaço de busca do problema de otimização que estamos interessados. Se chamamos a nossa métrica de qualidade de modelo funcional de c, transformamos nosso problema em

uma instância do problema de seleção de características. Assim, após definida a função de custo c e devidamente coletado e armazenado o conjunto S, propomos resolver instâncias do problema de identificação de vias de sinalização celular no arcabouço *featsel* [3].

3 Atividades Realizadas

3.1 Disciplinas cursadas

Durante o período de março a dezembro de 2018, foram cursadas pelo beneficiários três disciplinas. No primeiro semestre,

- Tópicos em Análise de Algoritmos: nesta disciplinas são abordadas técnicas para análise de algoritmos e para solução de problemas. Muitos problemas abordados nesta disciplina são de otimização combinatória, assim como o problema de seleção de características, que propomos utilizar neste projeto.

- Probabilidade e Inferência Estatística I: esta disciplina faz parte do departamento de Estatística do Instituto de Matemática e Estatística (IME-USP). A disciplina é dividida em dois módulos: o primeiro módulo aborda probabilidade, enquanto o segundo aborda inferência estatística, tanto do ponto de vista clássico quanto do ponto de vista Bayesiano. Esta disciplina foi necessária para que o beneficiário pudesse entender as duas funções de custo propostas para este trabalho: *Akaike's Information Criterion*, uma abordagem clássica; e *Bayesian inference-based modeling* (BIBm) [4], uma abordagem Bayesiana.

No segundo semestre, apenas a disciplina Laboratório de Programação Extrema foi cursada. Nesta disciplina, projetos com clientes reais são desenvolvidos pelos alunos usando a metodologia de programação extrema, uma metodologia ágil para desenvolvimento de sistemas. Algumas das técnicas ensinadas na disciplina estão sendo usadas no desenvolvimento

deste projeto, como controle de versões, desenvolvimento orientado a testes e integração contínua.

Além disso, no segundo semestre, o beneficiário frequentou como ouvinte a disciplina Introdução à Transdução de Sinais Celulares, no Instituto de Química da Universidade de São Paulo. Frequentando esta disciplina, o beneficiário pode se familiarizar com os processos bioquímicos envolvidos em uma via de sinalização celular.

3.2 Estudo de literatura em seleção de modelos

Apesar de propormos iniciar o projeto pela criação do banco de dados necessário para armazenar interações químicas relevantes assim como constantes de velocidades, decidimos iniciar o projeto pelo primeiro desafio científico reconhecido na proposta do projeto: definir uma função de custo apropriada para a seleção de modelos. Acreditamos que esta mudança tenha sido benéfica pois a criação do banco de dados necessitaria de um maior entendimento, pelo beneficiário, dos processos bioquímicos modelados. Assim, o beneficiário foi capaz de progredir na implementação de uma função de custo ao mesmo tempo em que se familiarizava com os experimentos biológicos que seriam modelados. Essa familiarização se deu pela participação do beneficiário em seminários promovidos pelo laboratório Especial de Toxinologia Aplicada, no Instituto Butantan, assim como sua participação como ouvinte na disciplina Introdução à Transdução de Sinais Celulares, no Instituto de Química da USP.

Assim, iniciamos o desenvolvimento deste projeto estudando funções de custo para modelos funcionais. Mais especificamente, decidimos começar estudando o *Bayesian inference-based modeling* (BIBm) [4]. De maneira superficial, nesta metodologia a qualidade de um modelo funcional é medida pela estimativa da probabilidade dos dados do experimento serem observados dado que o modelo em questão gera os dados observados; isto é, assumindo que o modelo avaliado representa bem o experimento biológico, medimos a probabilidade das observações feitas no experimento. Mais formalmente, dado um conjunto de observações experimentais D e um modelo M, a métrica aplicada no BIBm é o valor de $p(D|M)$.

Por ser uma abordagem Bayesiana, esta metodologia considera que o vetor de parâmetros de um modelo M é um vetor aleatório θ_M, de um espaço paramétrico Θ_M. Portanto, podemos escrever

$$p(D|M) = \int_{\Theta_M} p(D|M, \theta_M) p(\theta_M|M) d\theta_M \tag{1}$$

Ou seja, a métrica $p(D|M)$ é obtida ao marginalizar a verossimilhança $p(D|M, \theta_M)$. Entretanto, como estamos trabalhando com sistemas de EDOs, a verossimilhança muitas vezes não pode ser determinada analiticamente, assim como a integral que resulta na verossimilhança marginal $p(D|M)$. Para enfrentar este problema, a metodologia aplicada pelo BIBm usa como estratégia funções intermediárias entre a priori, $p(\theta_M|M)$, e a posteriori, $p(D|\theta_M, M)$, para obter um estimador de $p(D|M)$.

3.3 Estudos de estimadores de verossimilhança marginal de um modelo

Para entender como estimar a posteriori marginal $p(D|M)$, precisamos estudar os trabalhos de Friel e Pettitt [5], que utilizaram técnicas de integração de termodinâmica para calcular a integral 1, introduzindo um parâmetro de temperatura t que permite definir funções de probabilidade chamadas potência de posteriori, que são funções intermediárias entre a priori e a posteriori.

Note que os cálculos que faremos nesta seção são para um modelo M, portanto trabalharemos com probabilidades condicionadas ao modelo. Assim como Friel e Pettit (2008), vamos simplificar a notação desta seção ao remover das fórmulas o modelo M a qual estão condicionadas as probabilidades.

Define-se assim a função de probabilidade potência de posteriori:

$$p_t(\theta|D) = \frac{p(D|\theta)^t p(\theta)}{z(D|t)} \tag{2}$$

Com $z(D|t) = \int_\Theta p(D|\theta)^t p(\theta) d\theta$. Observe que $p_0(\theta|D) = p(\theta)$ é exatamente a distribuição a

priori dos parâmetros, enquanto $p_1(\theta|D) = p(\theta|D)$ é exatamente a distribuição a posteriori dos parâmetros, portanto, as distribuições potência de posteriori são capazes de traçar um "caminho" entre a priori e a posteriori quando se varia o parâmetro t de 0 a 1.

Agora considere $\frac{d}{dt}\log\{p(D|\theta)\}$. É possível mostrar que

$$\frac{d}{dt}\log\{z(D|\theta)\} = \int_{\Theta} \frac{p(D|\theta)^t p(\theta)}{z(D|t)} \log\{p(D|\theta)\} d\theta = \mathbb{E}_{\theta|D,t}[\log\{p(D|\theta)\}] \tag{3}$$

Além disso, é fácil ver que $z(D|t=0) = 1$ e $z(D|t=1) = p(D)$, a verossimilhança marginal que queremos estimar (lembre-se que estamos omitindo o condicionamento em M, o que significa que $p(D)$ é, na verdade, $p(D|M)$). Com estas informações, podemos escrever a identidade:

$$\int_0^1 \mathbb{E}_{\theta|D,t}[\log\{p(D|\theta)\}] = \log\{z(D|t=1)\} - \log\{z(D|t=0)\} = \log\{p(D)\} \tag{4}$$

A equação 4 nos permite estimar $\log p(D)$ sem calcular explicitamente a integral 1. Utilizaremos o mesmo estimador usado por Xu [4]. Primeiro, dividimos o intervalo $[0, 1]$ em $N-1$ pontos $r_1, ..., r_{N-1}$ a fim de criar N intervalos $T_1, ..., T_N$ tais que $T_i = r_i - r_{i-1}$, com $r_0 = 0$ e $r_n = 1$. Depois, para cada intervalo T_i, escolhe-se uniformemente neste intervalo N_i pontos $t_1^i, ..., t_{N_i}^i$. Finalmente, para cada uma das temperaturas t_j^i, amostra-se $\theta_j^i \sim \theta|D, t_j^i$. Assim, o estimador é dado por:

$$\hat{L} = \sum_{i=1}^{N} \frac{|T_i|}{N_i} \sum_{j=1}^{N_i} \log p(D|\theta_j^i) \tag{5}$$

Para obter esta amostras de $\theta|D, t_j^i$, Friel e Pettit utilizam um algoritmo chamado de *populational Markov chain Monte Carlo (MCMC)* (o chamaremos de MCMC populacional). Este algoritmo é similar ao algoritmo de Metropolis-Hastings [6] e é capaz de gerar simultaneamente amostras para todos valores de t_j^i. O valor de $p(D|\theta)$, como veremos na próxima seção, pode ser facilmente calculado quando definimos uma função de verossimilhança conveniente.

3.4 Implementação local do BIBm

Implementamos a metodologia do BIBm em um software que chamamos de SigNetMS (*Signaling Network Model Selection*). A linguagem escolhida para codificar este programa foi Python, devido ao grande número de bibliotecas e grande comunidade. O código é aberto sob a licença de software *GNU General Public License* e pode ser acessado em um repositório público do GitHub.

Este software recebe como entrada três arquivos diferentes: um arquivo SBML com a topologia do modelo funcional e concentrações iniciais de espécies químicas, determinando um modelo M; um arquivo XML que define a distribuição priori dos parâmetros do modelo M; e uma lista de arquivos que determina as observações D dos experimentos biológicos. O arquivo de experimentos deve especificar qual é a medida das observações, sendo esta uma função das concentrações das espécies químicas. Como saída, o software entrega uma estimativa de $p(D|M)$, o que chamamos de verossimilhança marginal de D.

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] 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.

[4] 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.

[5] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(3):589–607, 2008.

[6] John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin Andrew Gelman. *Bayesian Data Analysis, 3Rd Edn.* T&F/Crc Press, 2014.