Relatório Científico Final – 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 10 de dezembro de 2018 e 31 de dezembro de 2019

Laboratório Especial de Toxinologia Aplicada, Instituto Butantan

São Paulo, 14 de Janeiro de 2020

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 Atividades desenvolvidas

2.1 Disciplinas cursadas

2.2 Resumo de atividades anteriores

2.2.1 Estudo de seleção de modelos de via de sinalização celular

O desenvolvimento deste projeto se iniciou com o estudo de métodos capazes de avaliar a qualidade de um modelo de via de sinalização celular. Dado um conjunto de experimentos $\mathbf{D}$, que medem concentrações de espécies químicas, precisamos escolher uma função de custo $c(\mathbf{D}, M)$ que possa indicar a capacidade de um modelo M em reproduzir corretamente dados observados $\mathbf{D}$. O modelos de via que utilizamos é definido por um conjunto de reações químicas, produzindo um sistema de equações diferenciais, capaz de simular a dinâmica das concentrações de espécies químicas da via ao longo do tempo. Este sistema de equações diferenciais é criado utilizando leis de cinética química, como no modelo de Michaelis-Menten, e possuem constantes de velocidade que são usualmente desconhecidas; estas constantes são parâmetros dos modelos de vias.

A função de custo escolhida deve considerar possíveis valores para as constantes de velocidade do modelo avaliado. A abordagem de Lulu Wu [**?**], por exemplo, utiliza um processo de simulated annealing para encontrar o melhor conjunto de valores de parâmetros para um modelo e conjunto de experimentos. Entretanto, esta abordagem teve limitações que podem estar associadas a falta de informação a priori sobre as constantes e também a falta de penalização apropriada a modelos mais complexos. Por conta destas limitações, decidimos implementar uma função de custo baseada em estatística Bayesiana, chamada de verossimilhança marginal; denotamos $p(\mathbf{D}|M)$ a verossimilhança marginal de um conjunto de dados $\mathbf{D}$ dado um modelo M. Esta abordagem, apresentada no mesmo contexto no trabalho de Vyshemirsky e Girolami [**?**], permite a definição de informações a priori sobre constantes de velocidades e também induzem a penalização automática de modelos mais complexos.

Para calcular a verossimilhança marginal, precisamos definir a função de verossimilhança, $p(\mathbf{D}|M,\theta)$, onde $\mathbf{D}$ é o conjunto de experimentos, M é o modelo de interesse, e θ é um conjunto de valores para os parâmetros (constantes de velocidade) do modelo. Seguindo a abordagem de Kolch e Girolami, assumimos erro Gaussiano e independente:

$$p(\mathbf{D}|M,\theta) = \prod_{i=1}^{m} p_{\mathcal{N}_{(0,\sigma^2)}}([\phi(M,\theta) - \mathbf{D}]_i). \tag{1}$$

Onde $\phi(M,\theta)$ é um vetor com os valores simulados de concentrações, em cada intervalo de tempo, pelo modelo M usando parâmetros θ. A partir da função de verossimilhança, podemos obter a verossimilhança marginal com uma marginalização sobre os valores de parâmetros de modelos, ou seja, integrando a função de verossimilhança sobre o espaço paramétrico, Θ. Desta forma podemos escrever:

$$p(\mathbf{D}|M) = \int_{\Theta} p(\mathbf{D}|M,\theta)p(\theta|M)d\theta. \tag{2}$$

Entretanto, a integral 2 normalmente não pode ser calculada analiticamente. Para se calcular esta integral analiticamente, seria necessário determinar a distribuição de probabilidade conjunta $p(D,\theta|M)$, o que não é possível usualmente. Portanto, como é muito difícil (ou impossível) calcular a verossimilhança marginal, utilizamos um estimador desse valor como função de custo. Este estimador é construído utilizando um método conhecido como Integral Termodinâmica [**?**].

2.2.2 Estimação de verossimilhança marginal

O trabalho de Friel et al. [**?**] mostra que é possível reescrever o logaritmo da integral 2 como uma outra integral, um pouco menos simples, mas que nos permite criar estimadores para o logaritmo da verossimilhança marginal. Esta segunda forma de se escrever a verossimilhança marginal é baseada na integração de várias distribuições de probabilidade que são intermediárias entre as distribuições a priori e a posteriori das constantes de velocidade. As distribuições intermediárias são denominadas potências de posteriori.

Dada uma distribuição a priori $p(\theta|M)$ e a posteriori $p(\theta|\mathbf{D}, M)$, definimos a distribuição potência de posteriori como:

$$p_\beta(\theta) = \frac{p(\mathbf{D}|\theta, M)^\beta p(\theta|M)}{z(\beta)},$$

onde

$$z(\beta) = \int_\Theta p(\mathbf{D}|\theta, M)^\beta p(\theta|M) d\theta.$$

Note que $p_0(\theta)$ é a distribuição a priori e que $p_1(\theta)$ é a distribuição a posteriori. Portanto, podemos dizer que quando variamos o valor de β entre 0 e 1 estamos produzindo distribuição intermediárias que conectam a priori a posteriori. Friel et al. provam que é possível escrever:

$$\begin{aligned} \int_0^1 \mathbb{E}_{p_\beta(\theta)}[\ln p(\mathbf{D}|\theta, M)] d\beta &= \int_0^1 \frac{d}{d\beta} \ln z(\beta) d\beta \\ &= \Big[\ln z(\beta) \Big] \Big|_0^1 \\ &= \ln p(\mathbf{D}|M). \end{aligned} \tag{3}$$

O lado esquerdo da equação 3 recebe o nome de integral termodinâmica. Este nome se justifica pela variação do parâmetro β, que pode ser visto como um parâmetro de temperatura nas distribuições potência de posteriori. Esta integral pode ser estimada ou aproximada numericamente, permitindo acessar um valor próximo ao logaritmo da verossimilhança marginal. Para estimar ou aproximar esta integral, é necessário construir amostras de distribuições potência de posteriori para um conjunto finito de valores de β.

Em resumo, reescrevemos o logaritmo da verossimilhança marginal como uma integral que chamamos de integral termodinâmica. Esta integral pode ser aproximada numericamente ou estimada. Para ambas opções, é necessário escolher uma sequência de valores para β entre 0 e 1, e gerar amostras das distribuições potência de posteriori para os respectivos valores de β escolhidos.

2.2.3 Implementação do pacote SigNetMS

Após nossos estudos sobre seleção de modelos, decidimos implementar um pacote Python que nos providenciaria uma aproximação do logaritmo da verossimilhança marginal, usando os conceitos de integral termodinâmica. Este pacote foi implementado e recebeu o nome SigNetMS, e está disponível em um repositório público no GitHub [1]. O pacote SigNetMS recebe como entrada um arquivo no formato *Systems Biology Makup Language* (SBML) [**?**], com a definição das reações e constantes de velocidade da via; um arquivo *Extensible Markup Language* (XML) com resultados de experimentos; e um arquivo XML com definições de distribuições a priori para constantes de velocidade das reações da via. O pacote pode devolver como resposta o valor aproximado de $\log p(\mathbf{D}|M)$ e também amostras das distribuições potência de posteriori.

O pacote SigNetMS é capaz de processar modelos no formato SBML e criar os correspondentes sistemas de equações diferenciais ordinárias. Utilizando o pacote `Scipy` e seu integrador `odeint` é possível integrar esses sistema de equações diferenciais, criando uma simulação da dinâmica das concentrações gerada pelo par modelo e parâmetros (M, θ). Esta simulação é utilizada na função de verossimilhança, implementada de acordo com a equação 1. A verossimilhança do experimento, dado um modelo e conjunto de parâmetros, é usada no processo de geração da amostra de distribuições potência de posteriori e também na aproximação da verossimilhança marginal.

Para calcular a verossimilhança marginal, o pacote SigNetMS segue uma abordagem que faz uma aproximação numérica da integral 3. Esta aproximação é simplesmente a aplicação da regra dos trapézios para integrais. Desta maneira, é necessário escolher uma sequência de valores para β, o que também determina o conjunto de potências de posteriori que serão amostradas. O pacote SigNetMS faz esta escolha de $\beta_1, \ldots, \beta_T$ da maneira recomendada por Friel et al.:

$$\beta_t = \left(\frac{t-1}{T-1}\right)^c,$$

[1] https://github.com/gustavoem/SigNetMS

com $T = 20$ e $c = 5$. Assim, aplicando a regra dos trapézios na integral 3, podemos escrever:

$$\log p(\mathbf{D}|M) \approx \sum_{t=0}^{T-1} (\beta_{t+1} - \beta_t) \frac{\mathbb{E}_{p_{\beta_{t+1}}(\theta)}[\log p(D|M,\theta)] + \mathbb{E}_{p_{\beta_t}(\theta)}[\log p(D|M,\theta)]}{2}$$

Além disso, se considerarmos que a potência de posteriori β_t tem M_t parâmetros amostrados, então podemos substituir a esperança por um estimador de seu valor, produzindo a equação:

$$\log p(D|M) \approx \sum_{t=0}^{T-1} (\beta_{t+1} - \beta_t) \frac{\frac{1}{M_{t+1}} \sum_{i=1}^{M_{t+1}} \log p(D|M, \theta^{(t+1,i)}) + \frac{1}{M_t} \sum_{i=1}^{M_t} \log p(D|M, \theta^{(t,i)})}{2} \tag{4}$$

onde $\theta^{(j,i)}$ é o i-ésimo parâmetro amostrado para a potência de posteriori $p_{\beta_j}(\theta)$. Resta agora definir como as amostras de potência de posteriori são criadas.

As amostras de potência de posteriori são criadas em três etapas que utilizam o algoritmo Metropolis-Hastings. Esse algoritmo permite gerar uma amostra de uma distribuição (geralmente desconhecida ou difícil de se amostrar) a partir de uma distribuição de proposta, com a criação de uma cadeia de Markov. Chamamos estas três etapas de burn-in, burn-in informativo e amostragem final; todas estas etapas utilizam a distribuição log-normal como distribuição de proposta para os parâmetros.

Na versão do SigNetMS que utilizamos até a escrita do relatório parcial, a etapa de burn-in amostrava a distribuição a posteriori (ou seja, apenas uma cadeia) de parâmetros de maneira independente, com uma distribuição de pulo com covariância diagonal. Na etapa de burn-in informativo, uma amostragem similar a primeira etapa ocorria, porém utilizando uma distribuição de pulo com matriz de covariância diagonal tal que cada variância fosse igual a variância da amostra atual. Por fim, na ultima etapa, T cadeias eram geradas, uma para cada potência de posteriori escolhida, com distribuição de pulo igual a última utilizada na etapa anterior.

2.2.4 Primeiros testes da metodologia

Ainda no primeiro ano do projeto, testamos o SigNetMS na seleção de modelos. Porém, os resultados eram satisfatórios apenas para exemplos pequenos. Com exemplos maiores e com mais parâmetros, o pacote não apresentava bons resultados. Por esse motivo, começamos o segundo período do projeto (entre dezembro de 2018 e dezembro de 2019) ajustando nossa implementação

2.3 Melhorando a estimação da posteriori

Logo após a entrega do relatório parcial, identificamos que as amostras de potência de posteriori geradas pelo SigNetMS eram muito parecidas. Isso indicava que existia uma correlação grande entre as cadeias amostradas. Consultando o trabalho de Xu et al. [**?**] e Friel et al. [**?**] identificamos que as duas primeiras etapas de amostragem, burn-in e burn-in informativo também deveriam ser feitas para cada potência de posteriori escolhida, e não apenas para a distribuição a posteriori. Além disso, identificamos que a distribuição de pulo da etapa de burn-in informativo poderia ter como covariância a covariância amostral do conjunto de parâmetros aceitos até o instante.

2.4 Alternativa de avaliação de modelos

Ao mesmo tempo que investigamos possíveis erros na metodologia do SigNetMS também experimentamos uma outra função de custo Bayesiana para seleção de modelos de via, chamada ABC-SMC [**?**]. A função ABC-SMC se baseia em um método de geração de amostras conhecido como *Approximate Bayesian Computation* (ABC). Na função de custo ABC-SMC, amostras da distribuição $p(\theta, M|\mathbf{D})$ são geradas, permitindo estimar o valor de $p(M|\mathbf{D})$.

Podemos escrever um algoritmo genérico ABC que se propõe a gerar amostras da distribuição $p(\theta, M|\mathbf{D})$ com os seguintes passos:

1. Amostre um parâmetro candidato $(\theta^*|M^*)$ de uma distribuição de propostas.

2. Simule o par (M^*, θ^*) com os mesmos intervalos de tempo e para a mesma métrica do experimento $\mathbf{D}$, gerando $\mathbf{D}^*$.

3. Calcule, para alguma métrica de distância d, se o valor $d(\mathbf{D}, \mathbf{D}^*)$ for menor que um ϵ pré-determinado, então adicione o par (θ^*, M^*) a amostra.

4. Repita até uma condição de parada.

2.5 Paralelização do SigNetMS

2.6 Implementação eficiente de integração de sistemas de equações diferenciais

2.7 Testes da metodologia em uma cadeia do espaço de busca