Proposta do Trabalho de Conclusão de Curso

Aprendendo histórias: Aprendizado computacional para classificar histórias metabólicas

Larissa de Oliveira Penteado
Orientador: Roberto M. Cesar Jr

29 de Abril de 2017

Contextualização:

Ao estudar o metabolismo de organismos e processos celulares específicos, cientistas observam o conjunto de diferentes metabólitos presentes em uma célula, tecido, órgão ou organismo, conhecido também como metaboloma [1]. Quando expostos a diferentes condições fisiológicas, como alterações no meio em que vivem, os organismos apresentam mecanismos de resposta a tais estresses, que provocam mudanças nas concentrações desses metabólitos. Para compreender estas variações, existe o perfil metabolômico que representa os dados de concentração de metabólitos antes e depois de algum evento.

Uma rede metabólica representa um conjunto de reações químicas que regem o funcionamento do metabolismo de um organismo. Podemos modelá-la utilizando um grafo dirigido (digrafo), frequentemente chamado de grafo de compostos [2], ou um hipergrafo. Combinando um perfil metabolômico e a rede metabólica do organismo estudado é possível, através de métodos computacionais, inferir quais reações podem ter sido reguladas para produzir as mudanças de concentração observadas.

O problema de buscar uma explicação aos dados de metabolômica é conhecido na literatura pelo nome de *metabolite set enrichment analysis* [3]. Os métodos para *metabolite set enrichment analysis* utilizam vias metabólicas conhecidas na literatura. Porém ao estudar novos fenômenos, é interessante poder identificar vias alternati-

vas ou possíveis novas interpretações biológicas, ainda desconhecidas na literatura. Para isso pode-se utilizar o conceito de histórias metabólicas, que podem ser obtidas a partir de redes. [4, 5, 6, 7].

Neste projeto, lida-se com as histórias metabólicas obtidas através do método chamado Totoro, apresentado em Julien Laferrière(2016a; 2016b), em que as redes são modeladas como hipergrafos dirigidos, ou seja, os vértices são compostos e as hiper-arestas indicam as reações entre estes compostos. Eles são divididos em quatro categorias: os verdes, que representam os compostos cuja concentração aumentou; os vermelhos, que a concentração diminuiu; os brancos, nos quais ela se manteve inalterada; e os cinzas, cujas quantidades não foram medidas.

Uma história, nesse caso, é um conjunto mínimo de vértices e hiper-arestas que satisfaz as seguintes condições: todos os vértices verdes são ligados a vermelhos (ou seja, há uma hiper-aresta entre eles), e a recíproca também vale. Além disso, se vértices cinzas ou brancos fazem parte da solução, eles são substrato ou produto das reações. Vértices brancos, em especial, são sempre intermediários, ou seja, nunca são fonte ou alvo. Utilizando tal abordagem, não existe mais a necessidade de impor a condição de que o hipergrafo gerado pelo conjunto de hiper-arestas e vértices descrito acima seja acíclico. Este método pode ser acessado no site do Totoro.

Quando aplicado na rede metabólica da levedura, o método citado acima, produz aproximadamente 10^5 soluções. Como o número de histórias encontradas é muito grande, é preciso ter um modo de organizá-las, para que elas sejam melhor analisadas. Para tal, aplicarmos neste projeto a técnica de Aprendizagem (*Machine Learning*) para agrupá-las em categorias, e deste modo, poder interpretar as soluções obtidas. A análise e compreensão destas informações é importante, pois através disso podemos entender melhor como funcionam os mecanismos de resposta dos organismos, quando expostos a diferentes condições fisiológicas e estresses.

Principais atividades previstas para atingir os objetivos:

Na primeira etapa deste projeto, além dos estudos introdutórios ao tema, a meta é utilizar os dados do Madalinski et al (2008) , disponíveis no http://yeast.biocyc.org para download, para obter as redes metabólicas e utilizar o método descrito em Julien La-

ferrière(2016a; 2016b) para então gerar as histórias.

Em seguida, pretende-se tratar os dados obtidos acima (tanto as histórias como as redes metabólicas) para transformar suas representações em vetores de características que serão utilizados na etapa seguinte. Sejam $\mathbb{H}$ uma história metabólica encontrada pelo algoritmo Totoro e v o vetor de características que representa esta história. Se temos n reações na nossa rede metabólica, o vetor possui n posições, cada uma correspondendo a uma reação. Deste modo, se uma reação r está presente em $\mathbb{H}$, então $v[id(r)] = 1$, onde $id(r)$ é um número inteiro que representa a reação r, na rede metabólica. Caso contrário, $v[id(r)] = 0$.

Na terceira etapa, utilizaremos os vetores de características gerados na fase anterior para aplicar técnicas de Aprendizagem a fim de fazer uma classificação dos dados, ou seja, um *clustering*. Um dos métodos que será considerado nesta fase é o *K-means*, no qual queremos distribuir n amostras em uma partição de tamanho k, sabendo que uma amostra x pertence a uma parte T, se sua média é mais próxima à media da parte T do que das médias das outras partes. Além deste, pretendemos estudar e avaliar a aplicação de outras técnicas de aprendizado semi-supervisionado, cuja referência seria as vias metabólicas já conhecidas e descritas na literatura. Tais métodos serão avaliados no decorrer do projeto.

Cronograma:

- Março-Abril: Revisão bibliográfica do Totoro e demais leituras pertinentes ao projeto.
- Abril-Maio: *Script* para comparação de Métodos de Aprendizagem não-supervisionada de *clustering* da biblioteca *Python Scikit-Learn*.
- Maio-Junho: Escolha das métricas que serão utilizadas nos métodos de *clustering* e tratamento das histórias obtidas.
- Junho-Agosto: Aplicação do primeiro método de *clustering*, provavelmente o *K-means*. Início da composição da monografia.
- Agosto-Setembro: Escolha e aplicação de um segundo método de *clustering*.

- Setembro-Novembro: Análise dos resultados e composição do texto final.

Referências

[1] D. B. K. Stephen G. Oliver, Michael K. Winson and F. Baganz, "Systematic functional analysis of the yeast genome," *Trends in Biotechnology*, vol. 16, p. 373–378, 1998.

[2] L. Cottret, D. Wildridge, F. Vinson, M. P. Barrett, H. Charles, M.-F. Sagot, and F. Jourdan, "Metexplore: a web server to link metabolomic experiments and genome-scale metabolic networks," *Nucleic Acids Research*, vol. 38, no. suppl 2, pp. W132–W137, 2010. [Online]. Available: http://nar.oxfordjournals.org/content/38/suppl_2/W132.abstract

[3] J. Xia and D. S. Wishart, "MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data," *Nucleic Acids Research*, vol. 38, no. Web Server, pp. W71–W77, may 2010. [Online]. Available: http://dx.doi.org/10.1093/nar/gkq329

[4] V. Acuña, E. Birmelé, L. Cottret, P. Crescenzi, F. Jourdan, V. Lacroix, A. Marchetti-Spaccamela, A. Marino, P. V. Milreu, M.-F. Sagot, and L. Stougie, "Telling stories: Enumerating maximal directed acyclic graphs with a constrained set of sources and targets," *Theoretical Computer Science*, no. 457, pp. 1–9, 2012.

[5] P. V. Milreu, C. C. Klein, L. Cottret, V. Acuña, E. Birmelé, M. Borassi, C. Junot, A. Marchetti-Spaccamela, A. Marino, L. Stougie, F. Jourdan, P. Crescenzi, V. Lacroix, and M.-F. Sagot, "Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure," *Bioinformatics*, vol. 30, no. 1, pp. 61–70, 2014. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/30/1/61.abstract

[6] A. Julien-Laferrière, "Models and algorithms applied to metabolism: From revealing the responses to perturbations towards the design of microbial consortia ," Theses, Université

Lyon 1 - Claude Bernard, Dec. 2016. [Online]. Available: https://hal.inria.fr/tel-01394113

[7] A. Julien-Laferriere, "Totoro: Topological analysis of transient metabolic response," não publicada ainda, com previsão de publicação para o fim de 2016. [Online]. Available: http://hyperstories.gforge.inria.fr