Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Projeto e Análise de Algoritmos
1º Semestre de 2016

Recuperação de Redes Gênicas por meio de *Clustering* dos Perfis de Expressão dos Genes

Aluno: Geanderson Esteves dos Santos
geanderson@dcc.ufmg.br
Professores: Luiz Chaimowicz, Sebastián Urrutia, Wagner Meira Jr. e Jussara M. Almeida

1 Introdução

Este documento explica os principais conceitos relacionados ao problema NP-Completo escolhido como tema no trabalho final de Projeto e Análise de Algoritmos (PAA). O problema consiste na reconstrução de uma rede gênica reguladora a partir de dados de expressão gênica produzidos em larga escala pela tecnologia dos microarrays de DNA. A partir de uma série temporal de dados de várias amostras de expressão gênica, deve-se construir uma rede de relacionamentos entre genes que pode ser interpretada como uma rede gênica reguladora. Para cada um dos milhares de genes cuja expressão é amostrada em alguns instantes de tempo é gerada uma série de dados chamada de padrão ou perfil de expressão do gene no tempo [Akutsu et al. 1999] [D'haeseleer et al. 2000]. Estes perfis devem ser analisados a fim de se estabelecer relacionamentos entre os genes. Tais relacionamentos podem indicar que genes influenciam outros genes, isto é, participam do processo de regulação gênica [Davidson 2006]. Com o objetivo de clarificar a proposta deste trabalho, as próximas seções abordam alguns dos conceitos primordiais do trabalho.

1.1 Genes

Os genes [Alberts et al. 1989] são unidades informacionais básicas da hereditariedade. Eles representam sequências específicas de bases nucleotídicas, as quais carregam as informações necessárias para a construção das proteínas, que são as responsáveis pelos componentes estruturais das células, tecidos e enzimas dos seres vivos. Cada molécula de DNA é formada por diversos genes. O conjunto de todos os genes do DNA de um organismo é chamado de genoma.

1.2 Expressões Gênicas

A Figura 1 descreve de forma simplificada o processo de transcrição do material genético em RNA. Esse processo também é conhecido como expressão gênica, ou seja, um gene é considerado expresso toda vez que a proteína que ele codifica é sintetizada. A expressão gênica é responsável pelo processo de diferenciação das células dos seres vivos.

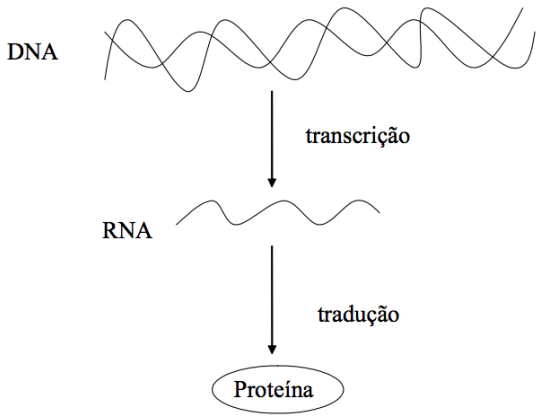

Figura 1: Processo de transcrição do DNA

1.3 Redes Gênicas

O comportamento dos genes nas células, i.e., quando eles devem ou não se expressar, é controlado pelas redes gênicas reguladoras, um mecanismo extremamente sofisticado capaz de interpretar os estímulos aos quais a célula está submetida [Bezerra 2006].

2 Modelagem em Grafos

A representação de redes gênicas reguladoras através de grafos foi proposta inicialmente por [Blais and Dynlacht 2005]. A Figura 2 mostra a modelagem proposta pelos autores. Cada vértice do grafo representará um gene, e os relacionamento entre os genes são as arestas do grafo.

3 NP-Completude

A extração de redes booleanas de dados de expressão gênicas em larga escala cresce exponencialmente, como mostrado por [Akutsu et al. 1999]. De forma complementar, o trabalho realizado por [Chen et al. 2001] confirma a np-completude do problema de inferir redes gênicas por meio de expressões gênicas.

4 Proposta do Trabalho

A proposta do trabalho é realizar a recuperação de redes gênicas por meio de agrupamento (*clustering*) dos perfis de expressão de genes, desde que estes representam a expressão de genes no tempo. A ideia é aplicar diferentes algoritmos de agrupamento e "cruzar" os dados resultantes. Os algoritmos podem ser K-means, EM (*Expectation Maximization*), disponíveis no ambiente WEKA [Witten et al. 2011] e algoritmos hierárquicos por aglomeração com esquemas *single-linkage* e *complete-linkage*[Jain et al. 1999].

Os algoritmos de agrupamento hierárquico permitem gerar facilmente estruturas de árvore, conhecidas como dendrogramas, que possibilitam a visualização entre grupos detectados nos dados e reconstruir a estrutura da rede.

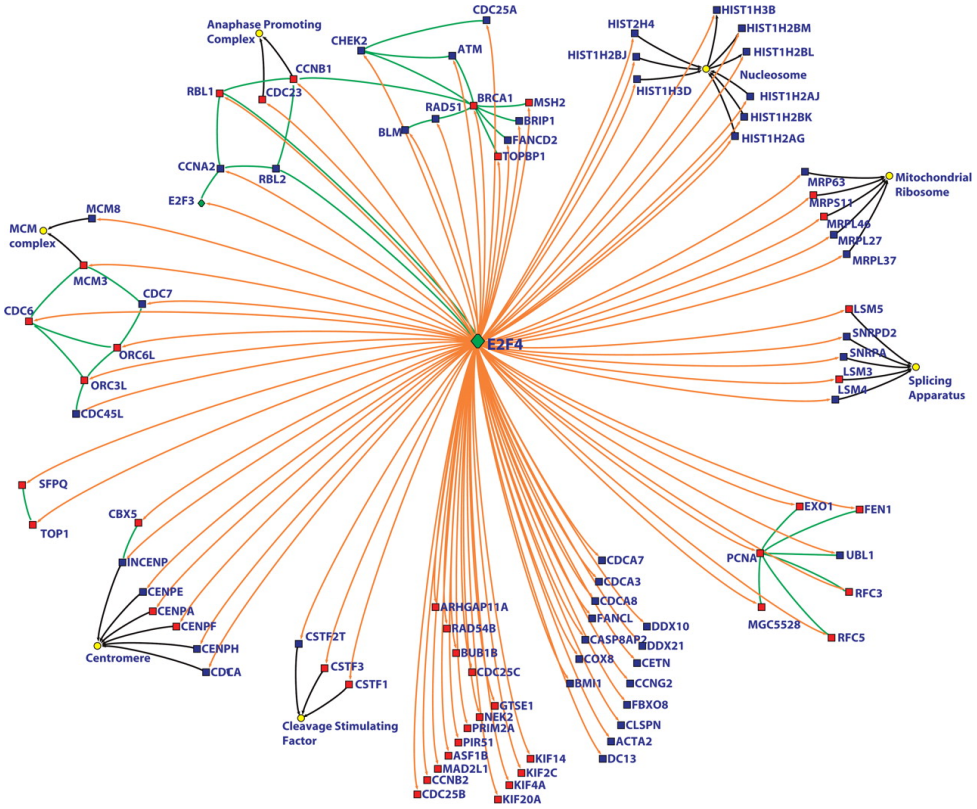

Figura 2: Redes gênicas reguladoras

4.1 Base de Dados

A base de dados definida para os testes do resultado da pesquisa foram gerados por [Tuomela et al. 2012]. Eles estão disponíveis no endereço (`ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE35nnn/GSE35103/matrix/`). A base de dados é interessante pois representa cerca de 700 genes para agrupamento no algoritmo de *clustering*.

Referências

[Akutsu et al. 1999] Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *US National Library of Medicine*.

[Alberts et al. 1989] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. (1989). *Molecular Biology of the Cell*. Garland.

[Bezerra 2006] Bezerra, G. B. P. (2006). Aplicações de Computação Bioinspirada em Bioinformática: Investigando o Papel dos Genes e suas Interações. Master's thesis, Universidade Estadual de Campinas.

[Blais and Dynlacht 2005] Blais, A. and Dynlacht, B. D. (2005). Constructing transcriptional regulatory networks. *Genes & Development*.

[Chen et al. 2001] Chen, T., Filkov, V., and Skiena, S. S. (2001). Identifying gene regulatory networks from experimental data. *Parallel Computing*.

[Davidson 2006] Davidson, E. H. (2006). *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution.* Academic Press, first edition.

[D'haeseleer et al. 2000] D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Oxford University Press.*

[Jain et al. 1999] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data Clustering: a review. *ACM Computing Surveys (CSUR).*

[Tuomela et al. 2012] Tuomela, S., Salo, V., Tripathi, S. K., Chen, Z., Laurila, K., Gupta, B., Äijö, T., Oikari, L., Stockinger, B., Lähdesmäki, H., and Lahesmaa, R. (2012). Identification of early gene expression changes during human Th17 cell differentiation. *American Society of Hematology.*

[Witten et al. 2011] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Series in Data Management Systems, third edition.