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

Reconstrução de uma Rede Gênica Reguladora por meio 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 tem o objetivo de explicar os conceitos relacionados a segunda entrega do trabalho final de Projeto e Análise de Algoritmos (PAA). Como requisitado pelo documento de entrega do trabalho, a modelagem em grafos do problema é apresentada de forma mais detalhada, alguns testes pertinentes do *baseline* são discutidos, e o plano de experimentos definido é mostrado. Além disso, outros conceitos relevantes e peculiares do problema de recuperação de redes gênicas por meio dos perfis de expressão dos genes. O *baseline* dessa pesquisa é o algoritmo de agrupamento K-means [Hartigan and Wong 1979] que foi aplicado pelos autores da principal referência desse artigo. Essa referência é importante, pois eles trabalham com uma base de dados interessante para análise de redes gênicas. É importante ressaltar que esses resultados são de extrema importância para especialistas em genética que usam bases de dados, tal como a proposta por [Tuomela et al. 2012], para estudar o combate de doenças em seres humanos. Saber o comportamento do gene, e como eles podem ser relacionados no tempo é de suma importância para esse tipo de pesquisa.

2 Baseline (Algoritmo de Agrupamento K-Means)

O algoritmo de agrupamento denominado *K-means* foi proposto por Hartigan em 1975, e melhorado pelo próprio autor anos depois [Hartigan and Wong 1979]. O objetivo do algoritmo é dividir um grupo de M pontos em N dimensões dentro de um número K de clusters. O valor de K determina em quantos *clusters* o algoritmo vai agrupar o conjunto de dados. De forma simplificada, o algoritmo agrupa um determinado conjunto de dados X em um espaço Y de conjuntos disjuntos. A Figura 1 mostra como o algoritmo funciona para um agrupamento de dois grupos (i.e., K = 2). O agrupamento é feito baseado em métricas de similaridade. Algumas das métricas mais utilizadas na literatura são: Distância Minkowsk, Distância Euclidiana, Distância Manhattan, Distância Mahalanobis, Métrica Overlap, Métrica VDM, Distância de Discriminação (DD) e Correlação de Pearson [de Carvalho 2003]. No entanto, a métrica de similaridade mais comum é a Distância Euclidiana, que tem sido utilizada comumente em plataformas de mineração de dados, como por exemplo, o weka [Witten et al. 2011].
Como exemplo de aplicação real do algoritmo, a pesquisa elaborada por [Tuomela et al. 2012] apresenta o agrupamento dos dados de 1373 genes considerados

diferencialmente expressos em cada instante de tempo, i.e., foram coletados dados nos instantes 0.5, 1, 2, 4, 6, 12, 24, 48 e 72 horas. Os autores aplicaram o algoritmo K-means, onde K é igual a 10, ou seja, 10 grupos de genes (i.e., subredes gênicas) criados a partir do conjunto de dados. Como a pesquisa de [Tuomela et al. 2012] é o *baseline* deste trabalho, o algoritmo K-means foi aplicado com o mesmo valor de K na linguagem de programação Java. O algoritmo se encontra em uma pasta denominada *src*. De forma complementar, o algoritmo pode ser acessado por qualquer computador através do GitHub no endereço (`https://github.com/gesteves91/KMeansArff`).
O arquivo denominado *clusters* contém os 10 agrupamentos criados pelo K-means dos 1373 genes.

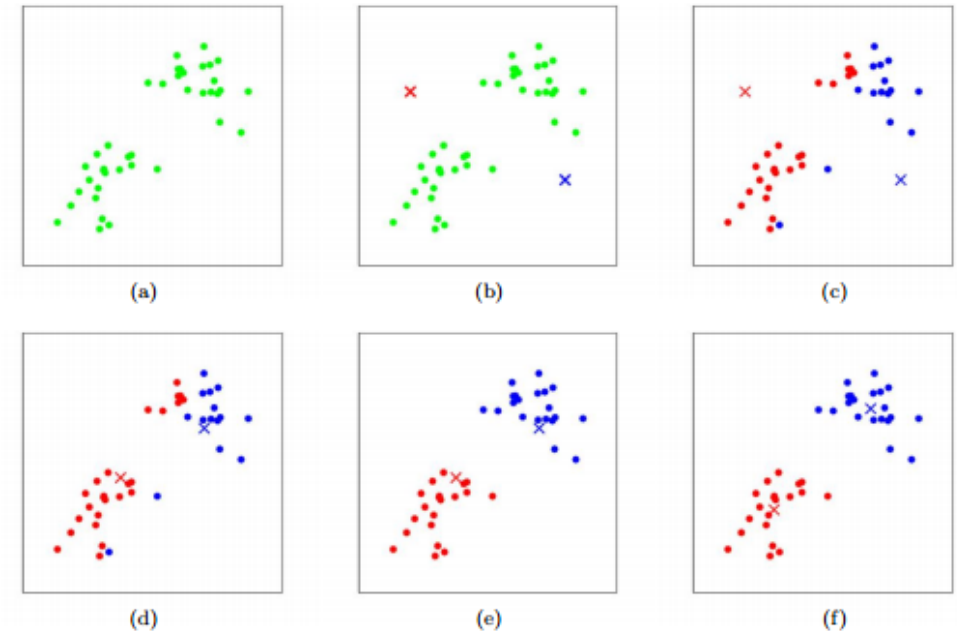

Figura 1: K-means em Funcionamento (k=2)

3 Modelagem em Grafos

A representação de redes gênicas reguladoras por meio de grafos foi mostrada por [Blais and Dynlacht 2005]. A Figura 2 mostra a modelagem proposta pelos autores. Cada vértice do grafo representará um gene, e os relacionamentos entre os genes são as arestas do grafo. A aplicação do K-means não necessariamente cria redes gênicas reguladoras, ou seja, redes nas quais os genes que têm alguma similaridade se regulam entre si, em contrapartida, o K-means apenas cria grupos de genes dentro de uma determinada rede. Para facilitar a análise que será realizada na parte final deste trabalho, foi desenvolvida uma função que cria um grafo baseado na relação que os genes têm entre eles. Cada grupo do K-means é um subgrafo completo de um grafo maior que contém todos os genes da rede (i.e., 1373 genes). O arquivo gerado pelo programa é denominado *graphs*, e armazena o grafo em uma matriz de adjacência (todos os arquivos explicados no documento estão no endereço do GitHub `https://github.com/gesteves91/KMeansArff`).

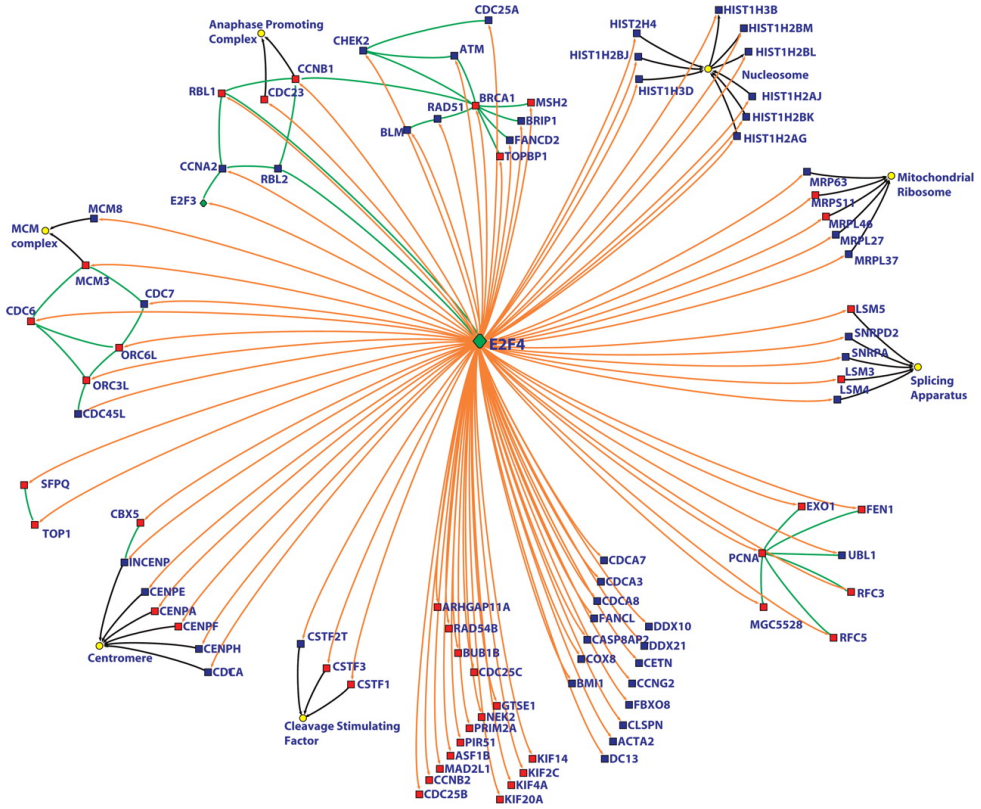

Figura 2: Redes gênicas reguladoras

4 NP-Completude

A extração de redes booleanas de dados de expressão gênicas é um problema NP-Completo, 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. O algoritmo K-means, apesar de seu comportamento polinomial, é apenas uma heurística para o problema de inferir redes gênicas reguladoras por meio de expressões gênicas. Não existe nenhuma garantia que os grupos criados pelo K-means representam uma rede gênica correta.

5 Solução Eficiente da Proposta

O algoritmo eficiente será implementado baseado em uma abordagem hierárquica de agrupamento com um paradigma que pode ser considerado guloso, tendo em vista que o melhor par de vértices (i.e., genes) serão escolhidos em cada etapa de execução do algoritmo. Esta solução será comparada com o *baseline* dos autores, ou seja, com a aplicação do K-means no trabalho de [Tuomela et al. 2012]. Entretanto a melhoria não deve ser em termos computacionais (i.e., tempo e/ou memória), mas em termos de:

- Riqueza ou quantidade de informações, ou seja, mais hipóteses para os especialistas analisarem;
- Qualidade das informações geradas, pois gerando-se subgrupos, teoricamente, detecta-se relacionamentos mais próximos entre genes e, provavelmente, mais

próximos da realidade e mais facilmente verificáveis, tendo em vista que são grupos menores.

Para a aplicação do K-means poderia ser argumentado que os subgrupos poderiam ser criados apenas aumentando o valor de K no algoritmo K-means, no entanto, aumentar K não gera subgrupos, necessariamente, apenas aumenta o número de grupos baseado na entrada de dados. Todavia, isso não significa que o algoritmo não pode ser computacionalmente mais rápido que o K-means. A complexidade do K-means pode ser descrita como:

$$O(nki) \tag{1}$$

Onde n corresponde ao número de dados (i.e, genes no contexto deste trabalho), k que é o número de *clusters* e i que é o número de iterações que o algoritmo executa, onde i é comumente definido como 500 devido a característica aleatória do K-means no momento de definição do centróide.

6 Plano de Experimentos Preliminares

A base de dados 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 ideia dos experimentos é comparar computacionalmente o K-means com o algoritmo hierárquico que será implementado, em termos de espaço e tempo, mas principalmente em termos dos grupos que ambos são capazes de criar. A hipótese que trabalhamos é de que o algoritmo hierárquico vai criar mais grupos levando em consideração outras estratégias de similaridades, todavia, como explicado na seção anterior, as maiores melhorias ocorreram em relação a qualidade dos dados produzida para análise de especialistas.

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*.

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

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

[de Carvalho 2003] de Carvalho, B. P. R. (2003). Métricas para cálculo de distâncias. Disciplina de Redes Neurais Artificiais (UFMG).

[Hartigan and Wong 1979] Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Algorithm AS 136*.

[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.