Problema da Clique Corrompida: Teoria de Grafos na Análise de Clusters de Genes[1]

Aimeé dos S. Reis* **Poly Hannah da Silva** **Simone Dantas**
Universidade Federal Fluminense
24020-140, Campus Valonguinho, Niterói, RJ
E-mail: aimeereis@id.uff.br / poly_hannah@id.uff.br / sdantas@im.uff.br

RESUMO

Os avanços da biotecnologia permitem pesquisas que medem o nível de expressão de milhões de genes simultaneamente nas diferentes condições e tempo. O *gene* é um segmento de uma molécula de DNA que contém um código de informações necessárias para a produção de proteínas. Analisar grandes quantidades de genes é uma tarefa difícil, por isso um dos métodos de pesquisa provém da análise e agrupamento dos genes que se manifestam com padrões de expressões similares.

Pequenos vetores medem o nível de expressão dos genes em *n* tempos diferentes. Esses dados são transformados nas *matrizes de intensidade* que permitem aos biólogos perceber como as funções dos genes podem ser relacionadas. Os dados são representados como pontos no espaço n-dimensional. Calculando-se a distância euclidiana entre cada dois genes, constrói-se uma *matriz de distância* dos genes (Figura 1). Genes com distâncias pequenas, os quais partilham as mesmas características e podem ser relacionados funcionalmente, são vistos na formação de *clusters* (grupos). Existem algumas técnicas de análise da formação de clusters, uma delas é baseada na teoria dos grafos. O modelo que estudamos a seguir foi proposto em [1].

Um *grafo G(V,E)* é um conjunto finito não-vazio V e um conjunto E de pares não-ordenados de elementos distintos de V. Uma *clique* de um grafo G*(V,E)* é um subconjunto S de V tal que *G[S]* é completo. Fixando um Θ que delimitará a distância entre os genes e utilizando a matriz de distância, podemos construir o *grafo de distâncias*: associamos um vértice para cada gene; e para cada par de genes, se a distância entre eles for menor que Θ, desenhamos uma aresta entre eles (Figura 1). Desse modo, no grafo de distâncias as cliques representam clusters. O *grafo de cliques* é um grafo onde cada componente é um grafo completo.

Um grafo pode ser transformado em um grafo de cliques, removendo-se ou incluindo-se arestas, conforme a Figura 2. Este é o chamado *Problema da Clique Corrompida*, cuja principal pergunta é: dado um grafo *G*, qual seria o menor número de arestas inseridas e removidas para transformar *G* num grafo de cliques? Este problema foi provado ser NP-Difícil e existem algumas Heurísticas para resolvê-lo [1].

Neste trabalho, estudamos o problema para as classes de grafos: caminhos e ciclos. Uma sequência de vértices $v_1,...,v_k$ tal que $(v_j, v_{j+1}) \in E, 1 \leq j \leq (k-1)$ é denominado *caminho* de v_1 a v_k,conforme a Figura 2. Um *ciclo* é um caminho $v_1,...,v_k$, sendo $v_1 = v_k$ e $k \geq 4$. Definimos como custo a soma do número de arestas removidas com o número de arestas inseridas necessárias para transformar o grafo em um grafo de cliques.

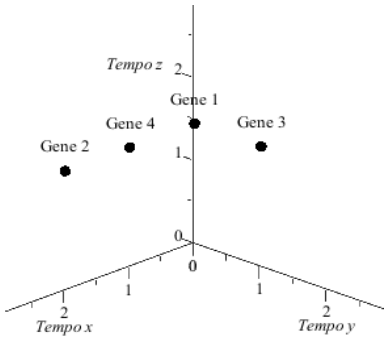
A técnica que nós desenvolvemos para a resolução do problema consiste em particionar o grafo formando grupos de dois vértices, e apenas um grupo de três vértices no caso ímpar. Comparamos nossos resultados com a heurística CAST [1], muito conhecida na literatura. Tal heurística particiona o grafo formando grupos de três vértices e apresenta para

* Bolsista de Iniciação Científica PIBIC/CNPq
[1]Financiado parcialmente por CNPq e FAPERJ

(a)

	Tempo X	Tempo Y	Tempo Z
Gene 1	1	1	2
Gene 2	3	1	2
Gene 3	1	2	2
Gene 4	2	1	2

(b)



(c)

	Gene 1	Gene 2	Gene 3	Gene 4
Gene 1	0	2	1	1
Gene 2	2	0	2,2	1
Gene 3	1	2,2	0	1,4
Gene 4	1	1	1,4	0

(d)

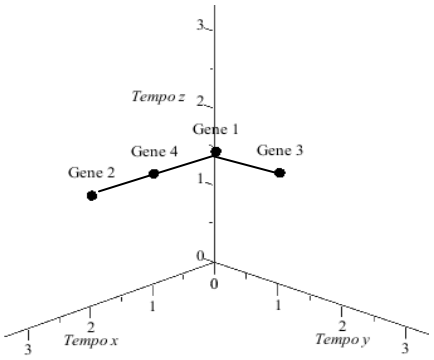


Figura 1: (a) Matriz de intensidade; (b) Gráfico de distâncias; (c) Matriz de distâncias; (d) Grafo de distâncias com Θ=1.

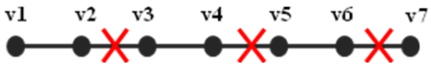


Figura 2: Caminho sendo transformado em cliques de tamanho 2.

caminhos um custo de $\frac{2n}{3} - 1$, se *n=3k*; e de $2\left\lfloor\frac{n}{3}\right\rfloor$, se *n=3k+1* ou $2\left\lfloor\frac{n}{3}\right\rfloor$, se *n=3k+2*. Para ciclos, apresenta um custo de $\frac{2n}{3}$, se *n=3k*; e de $\left\lfloor\frac{2n}{3}\right\rfloor + 1$, se *n=3k+1* ou $2\left\lfloor\frac{n}{3}\right\rfloor + 1$, se *n=3k+2*.

Descrevemos a seguir a nossa técnica desenvolvida. Para caminhos, se *n* for par, devemos retirar as arestas *(2i, 2i+1)*, $1 \leq i \leq \frac{n}{2} - 1$, tendo custo $\frac{n}{2}$ *-1*. Se *n* for ímpar, devemos retirar as arestas *(2i, 2i+1)*, $1 \leq i \leq \left\lfloor\frac{n}{2}\right\rfloor$ *-1* e incluir a aresta *(n-2, n)*, apresentando custo total $\left\lfloor\frac{n}{2}\right\rfloor$. No estudo dos ciclos, para todo *n* par, devemos remover a aresta *(1, n)* e as arestas *(2i, 2i+1)*, $1 \leq i \leq \left\lfloor\frac{n}{3}\right\rfloor$, tendo custo $\left\lfloor\frac{n}{3}\right\rfloor + 1$. Para todo *n* ímpar, devemos remover a aresta *(n, 1)*, as arestas *(2i,2i+1)*, $1 \leq i \leq \left\lfloor\frac{n}{3}\right\rfloor$ e inserir a aresta *(n-2,n)*, apresentando custo total $\left\lfloor\frac{n}{3}\right\rfloor + 2$. Desta forma, exibimos um custo menor que a heurística CAST. Por exemplo, considere um ciclo de *3858* vértices (*n=3858*), a heurística tem um custo total de *2572* e a nossa fórmula tem um custo total de *1287*, apresentando uma economia de *1285* arestas.

Palavras-chave: *Matemática Discreta, Teoria dos Grafos, Clique, Clusters de Genes.*

Referências

[1] A. Ben-Dor, R. Shamir, Z. Yakhini. Clustering Gene Expression Patterns, Journal of Computational Biology, vol. 6, pp. 281-297, (1999).

[2] J. A. Bondy and U. S.R. Murty, "Graph Theory with applications", University of Waterloo, Canada, 1976.

