O Transcriptogramer

João M. Dinis, Gabriel C. Perrone, Samoel R. M. da Silva, Rita M. C. de Almeida

Introdução

Para determiner o estado no qual o organismo se encontra devemos fazer uma medida quantitative de um RNA mensageiro, a forma mais usual de medi-la é com Microarranjo, uma medida quantitative de transcriptoma.

Durante a transcrição, expressão gênica, a informação genética do DNA é copiada em RNA mensageiro, em seguida é traduzida e finalmente codificáda em proteína, sendo transcriptoma o conjunto de todos os transcritos.

Transcriptograma

A medida de microarranjo é demasiadamente ruidosa, por isso foi desenvolvida uma técnica de tratamento de dados chamada Transcriptograma. Entre outras funções, essa técnica consiste em fazer uma média da expressão de genes relacionados para reduzir o ruído da medida utilizando uma *boxcar average*. Para isso ordenamos a lista de genes a partir de uma matriz de adjacência retirada do STRING, banco de dados de associação proteica, utilizando o algoritmo de metrópolis para minimizar a função E, aproximando proteínas associadas.

$$E = \sum_{i,j} a_{ij} D_{i,j}^{\alpha} I_{i,j}$$

Sendo D_{ij} a distância de a_{ij} até a diagonal pincipal da matriz e I_{ij} o termo de interface com suas respectivas expressões:

$$D_{i,j} = |i - j|$$

$$I_{ij} = |a_{i,j} - a_{i,j-1}| + |a_{i,j} - a_{i,j+1}| + |a_{i,j} - a_{i-1,j}| + |a_{i,j} - a_{i+1,j}|$$

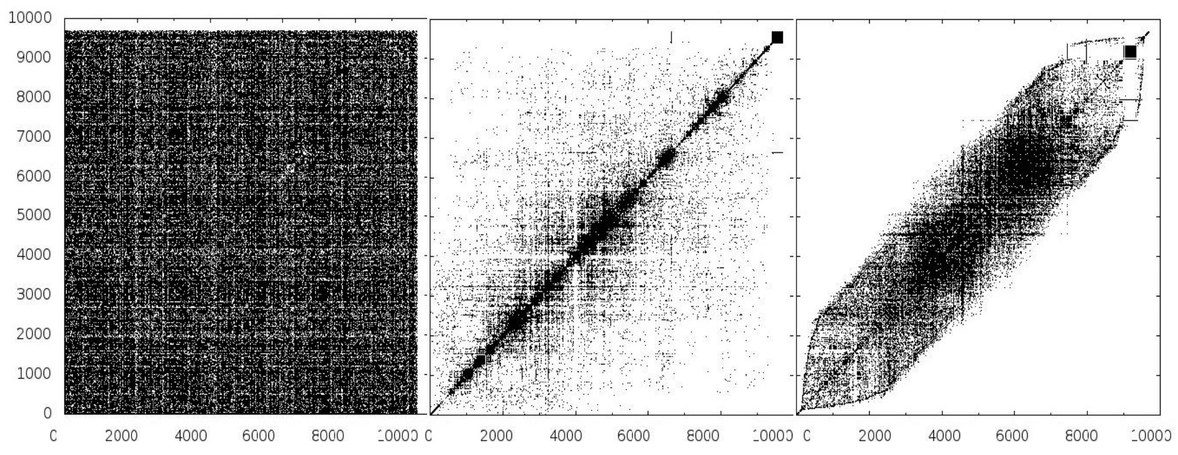

Figura 1: No painel acima, da direita para a esquerda, apresentamos a matriz de adjacência não ordenada, ordenada com $\alpha = 1$ e com $\alpha = 10$, respectivamente.

A partir da lista ordenada de proteínas, é possível fazer o transcriptograma, similar ao *Boxcar Average*, tomando a média da transcrição da vizinhança de cada ponto. O tamanho das vizinhanças é definido pelo raio r, como a média é feita com o gene e seus r vizinhos para a esquerda e para a direita o número total de proteínas avaliadas é $2r + 1$. Esta técnica além de reduzir o ruído também aumenta a reprodutibilidade das medidas.

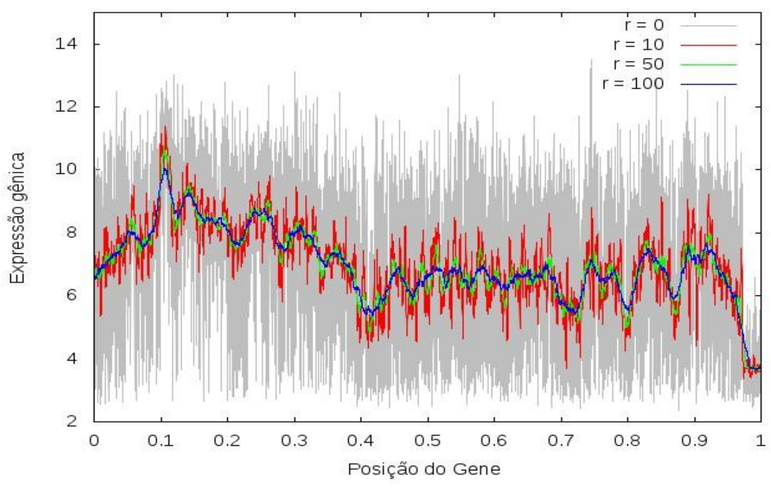

Figura 2: Diferentes raios de um transcriptograma para uma única amostra

Método de Diagnóstigo

Para realizar o diagnóstico, foram selecionados os melhores genes identificadores de cada classe a partir de um teste-t com variâncias desiguais realizado sobre as amostras de treinamento. A classificação das amostras foi feita com o método de aprendizado de máquina LDA (*linear discriminant analysis*).

A avaliação do diagnóstico foi feita com a métrica CCEM (*correct class enrichment metric*), que neste caso representa o percentual de acerto do diagnóstico

$$CCEM' = \sum_{i} p_i, c(i)\delta_i$$

onde, $p_{i,k}$ indica a confiança da previsão de i pertencer a classe k, $c(i)$ é a classe a qual a amostra i pertence e δ_i pode assumir dois valores, 1 se a amostra i foi corretamente classificada ou -1 se a amostra i foi incorretamente classificada.

Como o $CCEM'$ varia de valores de $-N$ a N normalizamos a expresão para obter uma eficiência de 0 a 1, dada por:

$$CCEM = \frac{\left(\frac{CCEM'}{2} + 1\right)}{2}$$

Resultados e Conclusões

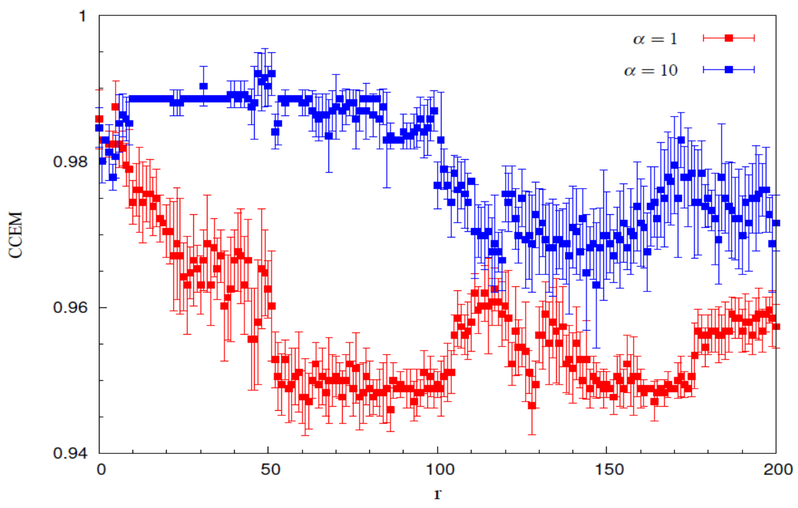

Figura 3: A eficiência do diagnóstico de psoríase em função do raio. Nota-se que para $r = 0$ estamos falando de dados de transcriptograma não tratados, ou seja, dados de transcriptoma.

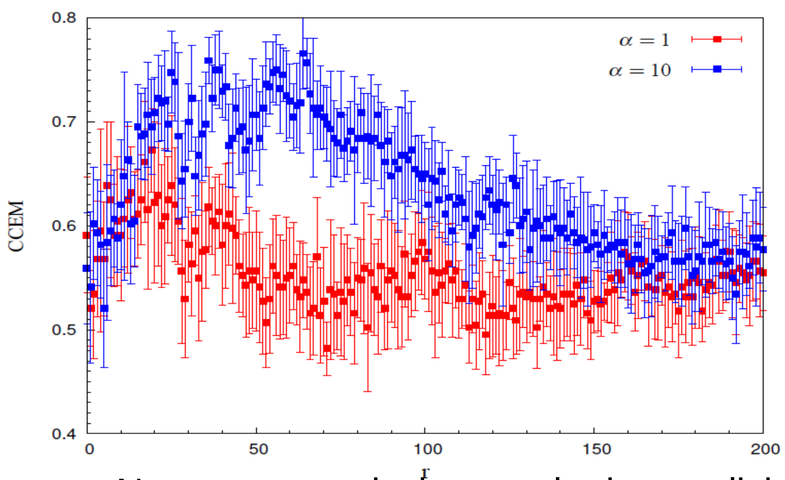

Figura 4: A eficiência do diagnóstico de psoríase em função do raio. Do mesmo modo que no gráfico anterior quando $r = 0$ estamos falando de dados de transcriptoma.

Nota-se, a partir dos resultados explicitados acima, que ocorre uma melhora significativa quando o diagnóstico utiliza-se de dados de transcriptograma.

A eficiência do método pode ser comprovada quando comparada aos resultados do time vencedor do desafio sbvIMPROVER, um programa de desafios focados em em pesquisas relacionadas a sistemas biológicos, na tabela abaixo.

Diagnóstico	Time vencedor	transcriptograma, $r = 0$	transcriptograma, eficiência máxima
Psoríase	0.9833	0.9847 ± 0.0038	0.9932 ± 0.0024
Esclerose múltipla	0.625199	0.5591 ± 0.0538	0.7659 ± 0.0340

Referências

[1] **Rybarczyk-Filho**, J.L., **Castro**, M.A.A., **Dalmolin**, R.J, **Moreira**, J.C.F., **Brunnet**, L.G. and **de Almeida**, R.M.C. ***Nucleic Acids Res.***, **39**, 3005-3016 (2011). PMID:**21169199**

[2] **Perrone**, G.C. **Master thesis**. 2013.Instituto de Física, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS,Brazil.

[3] **da Silva**, S.R.M. ***Master thesis***. 2013.Instituto de Física, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS,Brazil.