Universidade Federal do Rio Grande do Sul
Instituto de Física

A eficiência do transcriptograma

Samoel Renan Mello da Silva

Dissertação de mestrado

Porto Alegre, Setembro de 2013

Universidade Federal do Rio Grande do Sul
Instituto de Física
Programa de Pós Graduação em Física
Dissertação de Mestrado

A eficiência do transcriptograma[†]

Samoel Renan Mello da Silva

Dissertação de Mestrado realizada sob orientação da Prof. Rita Maria Cunha de Almeida, apresentada ao Instituto de Física, como requisito parcial para a obtenção do título de Mestre em Física.

Porto Alegre, Setembro de 2013

[†]Trabalho financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e pela Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS)

Resumo

A análise quantitativa do transcriptoma de um organismo revela informações quanto ao estado em que este se encontra, e uma técnica bastante usada para esta medida é o microarranjo, considerada bastante ruidosa. Pode-se perguntar se é possível usar as informações que possuímos acerca do funcionamento celular para o desenvolvimento de um método capaz de reduzir o ruído da medida. Usando a rede de interações entre proteínas de um organismo, desenvolveu-se a técnica chamada de transcriptograma [1], um método de ordenamento desta que permite a análise da expressão gênica em escala de genoma completo. Apesar de ter sido demonstrado que tal método permite a obtenção de resultados relevantes, inexiste até o momento análise mais criteriosa se a redução do ruído é verdadeira. Mostramos aqui que a medida é sim efetiva ao reduzir ruído intrínseco à técnica, e mais notavelmente é capaz de aumentar a confiabilidade da medida. Apesar de não podermos comparar o transcriptograma com nenhum tipo de padrão de ouro, podemos definir algumas estratégias para descobrir qual a eficiência de um transcriptograma indiretamente. Calculando o quanto de sinal que estamos atenuando ao reduzir o ruído, mostramos que o perfil de expressão produzido por um transcriptograma aparenta ser melhor para um conjunto de parâmetros diferente daquele usado em trabalhos anteriores. Aplicando o método a um problema de diagnóstico, mostramos que de fato a qualidade da informação obtida é maior para ordenamentos com estrutura e características diferentes do que se acreditava até então.

Abstract

Quantitative analysis of the transcriptome of an organism reveals information about its condition, and a widely used technique for this measure is the microarray, considered to be very noisy. One may ask whether it is possible to use the information we have about the cellular function for the method development to reduce the measurement noise. Using the network of interactions between proteins of an organism, it has been developed a technique called transcriptogram [1], an ordering method for this network that allows the gene expression analysis in whole genome scale. Although it was demonstrated that this method allows to obtain relevant results, a careful analysis to verify the method is still lacking. We show here that the method is effective at reducing the technique inherent noise and, most notably, is capable of increasing the the measurement reliability. Although we can not compare the transcriptogram with any gold standard, we can define some strategies to indirectly find out how efficient a transcriptogram is indirectly. Calculating how the signal is attenuated as noise is reduced, we show that the expression profile produced by a transcriptogram with the adequate set of parameters, is more efficient for diagnostic purposes, as compared to previous works. Furthermore, we propose in this work a method to obtain these optimal parameters. Finally, we apply the method to a diagnostic challenge and show that in fact the quality of the information obtained is higher for orderings with different structure and characteristics of what was believed until then.

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

Em um dos mais importantes acontecimentos científicos do século XX, a descoberta do DNA [7] teve um impacto imensurável nas áreas de medicina e biologia. Codificado em uma sequência dos nucleotídios guanina (G), adenina (A), timina (T) e citosina (C), chamadas de bases, o DNA contém toda a informação necessária para o desenvolvimento e funcionamento dos organismos vivos. Denomina-se de expressão gênica o processo no qual a informação contida em um gene é sintetizada em um produto gênico funcional que, em geral, são proteínas. Na transcrição, o primeiro passo da expressão gênica, a informação genética do DNA é copiado em RNA mensageiro, mRNA. No segundo passo, a tradução, o mRNA é por sua vez codificado em proteína.

O conjunto de todos os elementos transcritos em um célula é denominado transcriptoma, assim como proteoma é o conjunto das proteínas. Uma medida quantitativa de um RNA mensageiro é um indicativo do quão ativo está o gene correspondente e o quanto este está influenciando no comportamento celular, apesar deste não ser, efetivamente, o produto gênico funcional. Desta forma, uma medida quantitativa de transcriptoma, ou seja, a quantificação de todos os mRNA's presentes em uma célula em um determinado momento revela informações acerca do estado em que esta se encontra.

1.1 Microarranjo

Uma técnica de medida quantitativa de transcriptoma bastante utilizada é o microarranjo. O princípio de funcionamento está na atração química natural, através de pontes de hidrogênio, entre as bases que compõem o DNA: C forma um par com G, e A com T. O RNA é formado por três das bases formadores do DNA, A, G e C, mais uracil (U) no lugar de T, de modo que os pares de bases são formadas de forma equivalente ao DNA, C pareando com G, A com U. A ligação entre duas sequências de bases complementares é conhecido como hibridização.

O RNA, por ser formado por uma cadeia simples de bases, ao contrário da hélice dupla do DNA, irá se ligar com uma sequência complementar das bases que o formam,

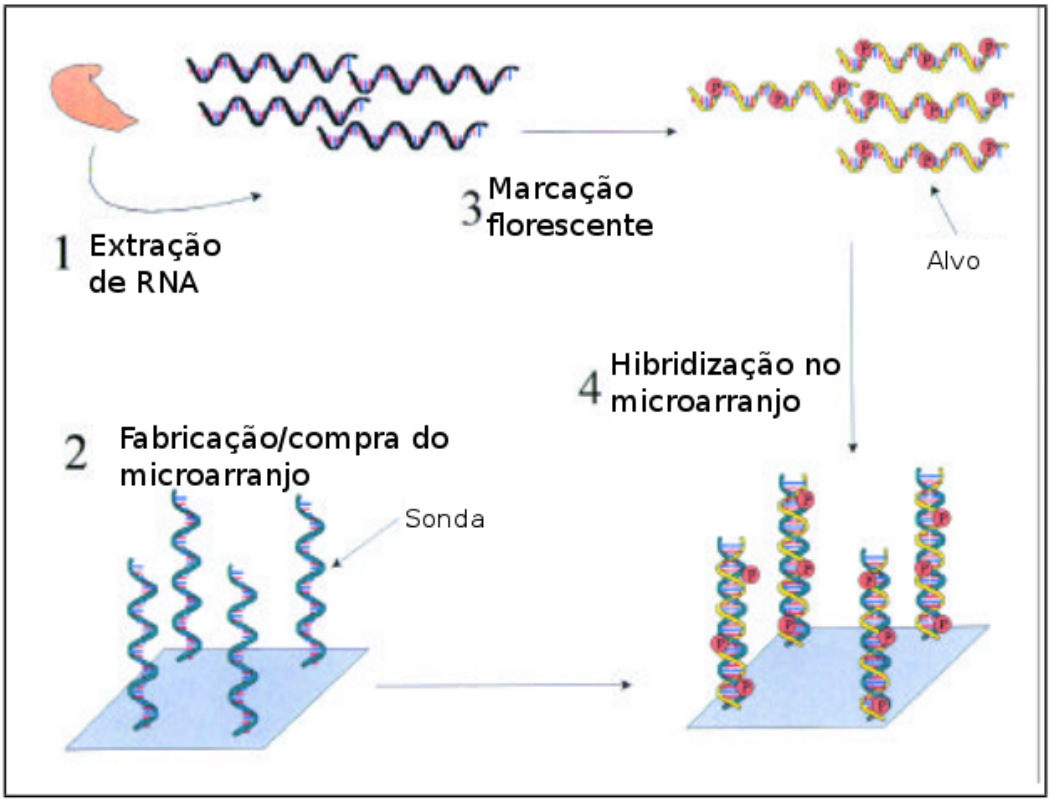


Figura 1.1: O RNA é extraído da amostra (1) e uma microarranjo com sondas específicas para aquele organismo é fabricado (2). A partir da amostra, obtêm-se alvos marcados com moléculas fluorescentes (3), que irão se hibridizar as suas sondas correspondentes. Adaptado de [2].

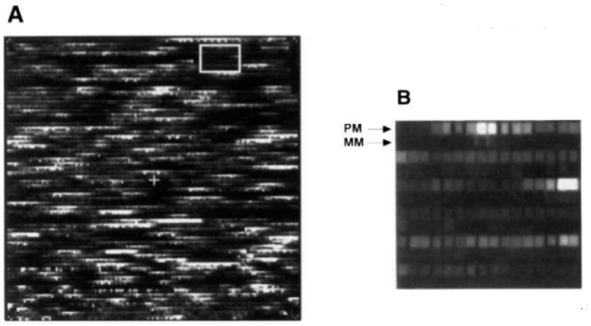


Figura 1.2: Imagem real de um microarranjo (adaptado de [3]): quanto maior a luminosidade de uma sonda, maior a quantidade do RNA, alvo desta, presente na amostra

seja ela DNA ou RNA. Em um microarranjo, uma sequência de pares de base é montada em cima de um *chip*, de modo a ler um dos RNA's presente na amostra (figura 1.1), que após marcação fluorescente permite sua quantificação (figura 1.2). Existem outras técnicas de medida quantitativa de expressão gênica, como o RNAseq [8] e o PCR [9], mas o microarranjo é o mais usado em medidas envolvendo o genoma completo.

Este trabalho se dedica a estudar uma técnica de análise de transcriptoma, o transcriptograma (Rybarczyk et al [1]) onde se usa uma rede de interações entre proteínas para montar uma lista ordenada de genes, de modo a aproximar aqueles que possuem correlação na sua expressão. Seria apressado discutir o método aqui, que se encontra explicado em detalhes mais adiante. A discussão que cabe no momento é como podemos responder às seguintes perguntas:

- dado que possuímos informações acerca do funcionamento celular em um organismo biológico, como podemos usá-la para melhor compreender a medida de transcriptoma? Sob certo aspecto, este trabalho pode ser entendido como um esforço para responder esta pergunta. Podemos, por exemplo, olhar para o mapa de rotas metabólicas montado pelo KEGG (enciclopédia de genes e genomas de kyoto), na figura 1.3, que apresenta mais dúvidas do que as soluciona.
- Ao se analisar a expressão do genoma completo, como podemos usar o conhecimento de interações bioquímicas para extrair informação da medida?

Esta dissertação está organizada de modo a, primeiramente, abordar o problema da medida e suas fontes de erro (capítulo 2), então apresentamos e caracterizamos o método de ordenamento de redes de proteínas e o ordenamento (capítulo 3), verificamos como os supostos problemas da medida se comportam quando tratado segundo o transcriptograma (capítulo 4) e verificamos a qualidade da informação que estamos extraindo (capítulo 5). Iremos nestes dois últimos aplicar o transcriptograma sobre importantes questões relativas a medida do transcriptoma por microarranjo. Em especial, discutiremos a reprodutibilidade da medida, ou seja, a comparação dos resultados de um mesmo experimento quando realizado por experimentos diferentes, e também o poder de diagnóstico que podemos realizar a partir desta técnica, onde os dados por esta produzidos são usados como única informação para a caracterização de amostras.

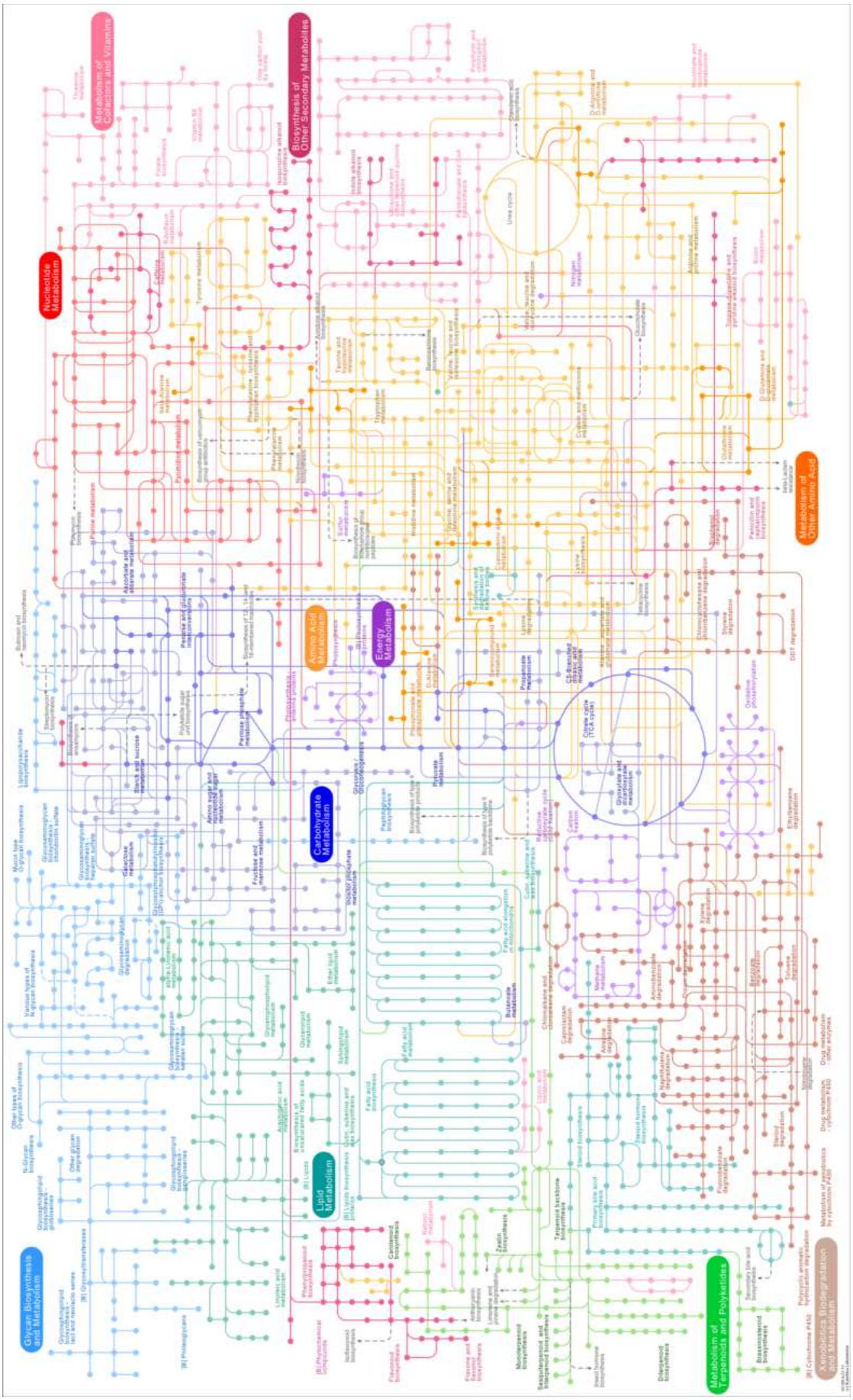


Figura 1.3: Mapa de rotas metabólicas do KEGG (Enciclopédia de Genes e Genoma de Kyoto) [4]), representando interações moleculares e rede de reações em uma célula.

2 Revisão

Este capítulo se dedica a descrever alguns aspectos que são importantes para o desenvolvimento deste trabalho. Inicialmente, vamos descrever os bancos de dados de onde provêm os dados que aqui foram utilizados. Ainda, será dada uma breve descrição de algumas propriedades de redes. Por fim, será explicado alguns aspectos do microarranjo, a fim de dar ao leitor maior conhecimento quanto ao problema básico ao qual este texto se dedica.

2.1 Bancos de dados

2.1.1 String

O STRING [10, 11, 12] (Search Tool for the Retrieval of Interacting Genes/Proteins), disponível em http://string-db.org, é uma base de dados que disponibiliza informações relativas a associações entre proteínas. Entenda-se por "associação" não só interação física direta entre proteínas, mas também interações indiretas, ditas funcionais, onde duas proteínas contribuem para a realização de determinada função sem que haja interação física entre estas. Ressalta-se ainda que o STRING tem como objetivo cobrir todo o espaço de associações possíveis, sendo provável que apenas um subconjunto destas se realize efetivamente em uma célula e/ou condição específica. É mantido e desenvolvido por um consórcio de instituições acadêmicas, listadas na tabela 2.1.

Tabela 2.1: Instituições desenvolvedoras do STRING

Sigla	**Nome**
CPR	*Novo Nordisk Foundation Center for Protein Research*
EMBL	*European Molecular Biology Laboratory*
KU	*University of Copenhagen*
SIB	*Swiss Institute of Bioinformatics*
TUD	*Dresden University of Technology*
UZH	*University of Zurich*

Para cada associação listada em sua base de dados, é disponibilizado um índice de

confiança, denominado *score s*, para cada evidência a partir da qual tal associação foi inferida ou observada. São sete as evidências listadas pelo STRING, que podem ser tanto agregações de informações de outros bancos de dados quanto critérios por este calculados. São elas:

- ***neighborhood***, ***fusion*** e ***co-occurrence***: estas três categorias têm por objetivo identificar pares de genes que parecem ter sofrido pressão seletiva comum ao longo de processos evolutivos e seriam, portanto, funcionalmente associados. Isto é feito a partir da comparação do genoma de vários organismos. Mais especificamente, o critério *neighborhood* é obtido a partir da proximidade genômica entre um par de proteínas em diferentes organismos, *fusion* é obtido para par de genes que sofreram fusão em outras espécies, enquanto *co-occurrence* é obtido por padrões de presença/ausência de genes em genomas de diferentes organismos;

- ***co-expression***: calculado a partir de padrões de coexpressão em experimentos de expressão genômica publicados na literatura;

- ***experimental*** e ***database***: aqui, informações de diversos bancos de dados são agregados, seja relativas a experimentos diretos de associação física (*experimental*) ou aqueles que descrevem rotas metabólicas, funções biológicas ou afins (*database*). Os bancos de dados importados estão listados na tabela 2.2.

Tabela 2.2: Bancos de dados importados pelo STRING

Sigla	**Nome**	**Disponível em**
BIND	*Biomolecular Interaction Network Database*	bind.ca
DIP	*Database of Interacting Proteins*	dip.doe-mbi.ucla.edu/
BioGRID	*Biological General Repository* for *Interaction Datasets*	thebiogrid.org/
HPRD	*Human Protein Reference Database*	www.hprd.org/
IntAct	*Intact*	www.ebi.ac.uk/intact/
MINT	*Molecular Interaction Database*	mint.bio.uniroma2.it/
PID	*Pathway Interaction Database*	pid.nci.nih.gov/
Biocarta	*Biocarta*	www.biocarta.com/
BioCyc	*BioCyc Database Collection*	www.biocyc.org/
GO	*Gene Ontology*	www.geneontology.org/
KEGG	*Kyoto Encyclopedia of Genes and Genomes*	www.genome.jp/kegg/
REACTOME	*Reactome*	www.reactome.org/

- ***textmining***: coocorrência de nomes de genes ou proteínas em resumos de artigos;

Para cada evidência de cada associação é disponibilizado um valor de confiança para a que a associação seja verdadeira, denominado *score*. Define-se como uma associação verdadeira a ocorrência de ambas as proteínas associadas na mesma rota metabólica do KEGG [4] (Enciclopédia de Genes e Genomas de Kyoto). O KEGG é considerado um bom padrão de ouro pois este cura manualmente seus dados e os mesmos cobrem grande variadade de organismos.

O *score* total para a associação é calculado conforme [12]

$$S = 1 - \prod_i (1 - S_i), \tag{2.1}$$

onde S_i indica o *score* individual de cada evidência.

O STRING se destaca como um banco de dados livre, de fácil acesso para obtenção de dados em larga escala e com grande quantidade de dados e informações, o que o torna adequado para este estudo. Neste trabalho, foram utilizados dados da versão String 9.05, que contém mais de 5 milhões de proteínas e 200 milhões de interações para mais de 1000 organismos. Foram consideradas associadas todo par de proteínas cujo score de associação fosse igual ou superior a 0.8, sendo consideradas não associadas caso contrário. Um exemplo de rede de interações entre proteínas que podemos retirar do STRING pode ser vista na figura 2.1.

2.1.2 Gene Ontology

O Gene Ontology [13], disponível em http://www.geneontology.org/, é uma base de dados que classifica genes e seus produtos com base em três critérios:

- Componente celular: diz respeito ao local na célula onde o produto gênico está ativo. Exemplos de ontologias de componente celular são “ribossomo” (GO:0005840) e “complexo de golgi” (GO:0005794);
- Função molecular: é a atividade bioquímica do produto gênico. Exemplos de funções moleculares são “atividade catalítica” (GO:0003824) e “atividade receptora” (GO:0004872);
- Processo biológico: se refere ao objetivo biológico para o qual o gene contribui, como “tradução” (GO:0006412) e “crescimento celular” (GO:0016049).

As ontologias são organizadas de forma estruturada, se assemelhando a uma hierarquia onde termos filhos são mais especializados e termos pais sendo mais generais. Por

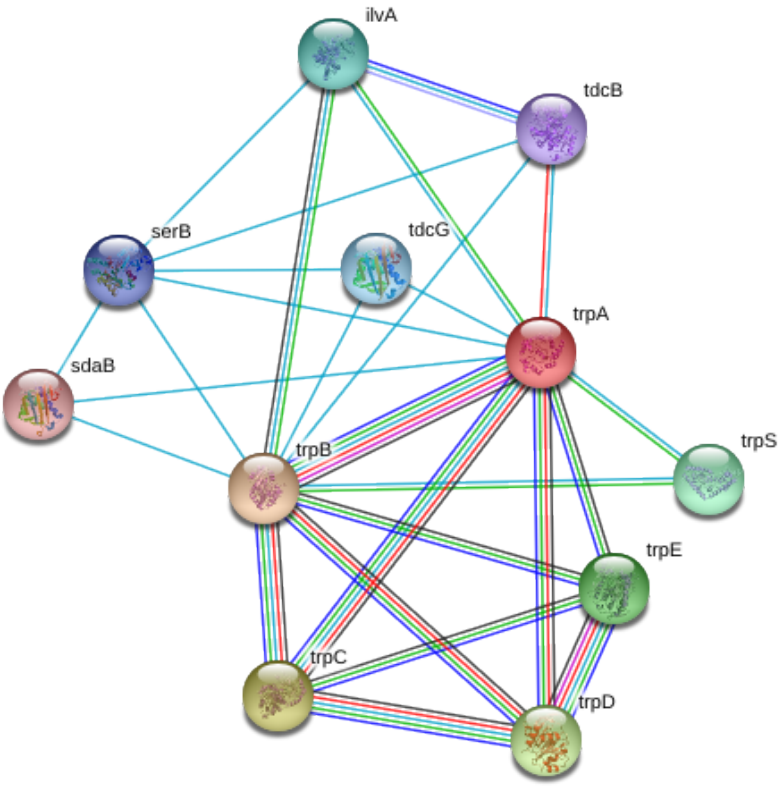


Figura 2.1: Exemplo de rede de proteínas, retirado do STRING, para algumas proteínas da *Escherichia coli*. Diferentes cores indicam diferentes fonte de evidência para existência da interação entre proteínas.

exemplo, a componente celular mitocôndria (GO:0005739) é um termo filho de organela (GO:0043226). No entanto, diferentemente de uma hierarquia, um termo dito filho pode ter mais de um termo pai. Um exemplo deste fato é o processo biológico processo biossintético de hexose (GO:0019319) que possui dois termos pais, processo metabólico de hexose (GO:0019318) e processo biossintético de monossacarídeo (GO:0046364), que ocorre porque um processo biossintético é um tipo de processo metabólico e hexose é um tipo de monossacarídeo.

2.1.3 Gene Expresion Omnibus e ArrayExpress

Estes dois bancos de dados, o Gene Expression Omnibus [14, 15] (GEO), disponível em http://www.ncbi.nlm.nih.gov/geo/, assim como o ArrayExpress [16], disponível em http://www.ebi.ac.uk/arrayexpress, são bases de dados que arquivam e distribuem livremente dados de experimentos de microarranjo, entre outros experimentos relativos a genômica. O GEO é mantido pelo Centro Nacional para Informação Biotecnológica (NCBI) dos Estados Unidos, enquanto O ArrayExpress, é parte do Instituto Europeu

de Bioinformática (EBI), uma divisão do Laboratório Europeu de Biologia Molecular (EMBL). Por regra geral, exige-se que, ao se publicar um artigo científico onde se analisam dados de transcriptoma por microarranjo, os dados sejam livremente disponibilizados em algum banco de dados. Destacam-s o GEO, atualmente com 40686 experimentos publicados, e o ArrayExpress, com 40868 experimentos.

2.1.4 Biomart

Disponível em http://www.biomart.org, o projeto BioMart se dedica a fornecer à comunidade científica internacional um conjunto de informações relativas aos vários bancos de dados existentes na área de biologia, com o objetivo de facilitar a integração dos dados fornecidos por estes. De fato, neste trabalho, faz-se necessário a integração das informações obtidas nos bancos de dados acima descritos, e a diferença de nomenclatura entre estes torna necessário que estes dados sejam traduzidos, por assim dizer. O Biomart configura uma ferramente adequada para tal, devido à facilidade da sua utilização e sua abrangência.

2.2 Microarranjo Affymetrix

Todos os experimentos de microarranjo que foram analisados aqui são de microarranjos fabricados pela empresa Affymetrix. Apesar de não ser a única fabricante de microarranjos, os produtos fabricados pela Affymetrix são aqueles mais utilizados atualmente, como se pode ver na tabela 2.3, que lista as plataformas mais utilizadas segundo o GEO. Não se exclui a validade do que é demostrado neste trabalho para microarranjos de outras fabricantes, visto que os princípios básicos da técnica se assemelham.

Tabela 2.3: Plataformas de microarranjo mais utilizadas, segundo pesquisa realizada junto ao GEO.

Plataforma	Fabricante	Amostras	Experimentos
HG-U133_Plus_2	Affymetrix	84738	3104
HG-U133A	Affymetrix	34060	965
HumanHT-12 V3.0	Illumina	15051	335

Foge ao escopo deste trabalho a investigação dos chamados métodos de pré-processamento dos dados de expressão oriundos de um microarranjo Affymetrix. Entretanto, se faz necessário aqui uma descrição breve de tais métodos, visto se tratar de procedimento necessário para a análise dos dados de expressão, e também para dar ao leitor uma maior compreensão sobre o funcionamento da técnica e das dificuldades envolvidas em sua análise. Ainda, cabe ressaltar que não raras vezes estes métodos são chamados de

"normalizações" na literatura, mas tal termo deve ser interpretado com cuidado. De fato, os métodos de pré-processamento dos dados geralmente contém uma etapa de normalização dos dados, como se verá a seguir, mas não são apenas isso.

Primeiramente, cabe explicar que em um microarranjo Affymetrix cada gene é representado não por uma sonda, mas sim por um conjunto de sondas, sendo cada conjunto composto por dezenas de pares de sonda. Cada par de sonda, por sua vez, consiste em uma sonda chamada PM (*perfect match*) e uma sonda MM (*mismatch*). Uma sonda PM contém uma sequência de 25 bases que corresponde exatamente a uma sequência do gene ao qual a sonda se destina a ler. Uma sonda MM, por sua vez, é idêntica à sonda PM com a qual faz par, mas a base do meio, a 13ª, é diferente. Assim, uma sonda MM não deveria ler, idealmente, sinal algum, visto que sua sequência é planejada para não ser complementar a nenhum RNA da amostra.

Antes de iniciarmos a descrição dos métodos de pré processamento, vamos definimos inicialmente os índices i, que denota a amostra ou microarranjo, j, que denota o conjunto de sondas destinado a ler determinado gene, e k, que denota um par de sonda específico contido em um conjunto de sondas, para identificar cada sonda PM e MM. Ainda, diga-se que estes métodos envolvem três etapas distintas: a correção de fundo, para retirar um sinal de fundo da medida como um todo, a normalização dos dados, e o tratamento do sinal da sonda PM em relação à sonda MM. Daremos maior atenção a esta última por ser a mais relevante para a compreensão da técnica.

2.2.1 AvDiff

Inicialmente, os valores de expressão eram calculados como a média das diferenças de intensidades lidas no par PM/MM, o que significa assumir que todo erro de medida daquela sonda PM, seja óptico ou biológico, é lido na sonda MM. De modo a evitar que valores inconsistentes alterem muito o valor de expressão, são excluídos da média os pares cuja diferença se distancie a mais de 3 desvios padrões da média:

$$(AvDiff)_{ij} = \frac{1}{N_A} \sum_{k \in A_{ij}} (PM_{ijk} - MM_{ijk}) \tag{2.2}$$

onde A é o subconjunto do conjunto de sondas j da amostra i no qual $\Delta_{ijk} = PM_{ijk} - MM_{ijk}$ não se distancia da média de Δ_{ijk} sobre os k-ésimos pares PM/MM por mais de 3 desvios padrões.

Observa-se que em um experimento típico de microarranjo, cerca de terço das sondas MM é mais intensa que seu par PM [17]. Apesar da tentativa de se livrar dos valores

inconsistentes, este método ainda produz valores de expressão negativos para alguns genes, o que é fisicamente impossível. A AvDiff caiu em desuso, dando lugar, principalmente, aos dois métodos descritos a seguir.

2.2.2 MAS5

Na correção de fundo, primeira etapa do método, o microarranjo é dividido em 16 regiões retangulares, e define-se como o sinal de fundo de cada região a média das sondas que estão entre as 2% menos expressas. A correção de fundo se dá então tendo como base sua posição física no microarranjo e o sinal de fundo calculado para cada região do microarranjo.

Aqui, ainda se mantém a hipótese de que o erro da medida é lido na sonda MM. Entretanto, nas sondas onde a intensidade lida na MM é maior que a PM, essa hipótese levaria à conclusão de que aquele valor de expressão é negativo, fisicamente impossível. Para contornar esse problema, considera-se que nestes pares a sonda MM falha ao ler o erro da medida, e este é então estimado a partir das demais sondas do conjunto de sondas a que este pertence. Define-se, portanto, o erro ideal, IM,

$$IM_{ij} = \begin{cases} MM_{ij}, & MM_{ij} < PM_{ij} \\ \frac{PM_{ij}}{2^{SB^{+}_{ij}}}, & MM_{ij} \geqslant PM_{ij} \end{cases}, \tag{2.3}$$

que será igual a sonda MM caso seu valor seja inferior a PM, ou uma fração do sinal da PM em função de um ruído específico positivo daquele conjunto de sondas j, $SB_i j^{+}$. Este ruído, por sua vez, é dado por

$$SB^{+}_{ij} = \begin{cases} SB_{ij}, & SB_{ij} > \tau \\ \frac{\tau}{1+0.1(\tau - SB_{ij})}, & SB_{ij} \leqslant \tau \end{cases}, \tag{2.4}$$

$$SB_{ik} = \mathrm{TB}_j \left(\log_2(PM_{ijk}) - \log_2(MM_{ijk}) \right), \tag{2.5}$$

onde TB_j significa o cálculo da média sobre o índice j usando *Tukey's Biweight* (apêndice A). E o parâmetro $\tau = 0.03$. A equação 2.4 diz que o ruído é calculado a partir do erro lido naquele conjunto de sondas, mas caso isso também falhe em produzir um erro maior que PM, SB^{+}_{k} segue (eq. 2.4, segundo caso) fracamente baseado nos dados daquele conjunto de sonda. Finalmente, o valor de expressão é corrigido com

$$PM'_{ijk} = max(PM_{ijk} - IM_{ijk}, 2^{-20}), \tag{2.6}$$

onde $max(a,b)$ indicia o maior valor entre a e b.

Finalmente, o valor de expressão é calculado por

$$PV_{ij} = TB_k(\log_2(PM'_{ijk})). \tag{2.7}$$

Por fim, na normalização, a última etapa do método, realiza-se uma normalização constante, onde todos os valores de expressão de são transladados por um determinado valor de modo que a expressão média de uma amostra seja igual ao de um valor alvo, por padrão, 500.

2.2.3 RMA

No método *Robust MultiArray* (RMA) [17], novamente a primeira etapa é o de retirar o sinal de fundo da medida. Aqui, supõe-se que a distribuição do sinal em relação a sua intensidade é a soma de um sinal verdadeiro, que decai exponencialmente, e um ruído com distribuição normal. Para a normalização, que aqui não é a última etapa, usa-se o método conhecido como normalização *quantile* [18], onde o objetivo é tornar as distribuições idênticas em propriedades estatísticas.

Para a terceira etapa, o RMA assume que as sondas MM não trazem informação confiável quanto ao erro medido em PM. Se em um terço dos casos a leitura da sonda MM é maior do que o da PM, argumenta-se que uma sonda MM também é capaz de ler sinal verdadeiro, não apenas o erro de medida. Deste modo, calcula-se o valor de expressão para o conjunto de sondas de cada gene baseado apenas nas sondas PM. Assume-se ainda que o erro de medida é multiplicativo e que o sinal lido é dependente de um termo de afinidade. De fato, observa-se que o sinal da sonda MM é tão maior quanto maior for o sinal da sonda PM [17], forte indício de que a sonda MM detecta sinal verdadeiro. Deste modo, seja Y_{ijk} o sinal observado em PM, após correção de fundo e normalização, em escala logarítmica este será dada por

$$Y_{ijk} = \mu_{ij} + \alpha_{jk} + \varepsilon_{ijk}, \tag{2.8}$$

onde μ_{ij} é o sinal do gene j na amostra i e α_{jk} é a afinidade da sonda k do gene j. O termo ε_{ijk} representa o ruído da medida. Note, o termo de afinidade da sonda α é o mesmo para toda amostra i, e o pré-processamento conjunto de todas as amostras de um mesmo experimento, para descobrir a afinidade de cada sonda, é um conceito chave que diferencia o RMA de outros métodos de pré-processamento. Após ajuste da equação 2.8 às expressões observadas nas sondas PM, μ_{ij} é o valor de expressão obtido pelo método RMA.

Podemos considerar uma questão em aberto sobre qual é o melhor método de pré-

processamento possível, havendo trabalhos que se dedicam especificamente a analisar qual produz melhor resultados [19] [20] [21], mas é consenso que o RMA supera outros métodos para genes com baixa expressão, onde o MAS5 produz muitos falsos positivos [17] [21]. Diga-se também que existem outros métodos, menos utilizados que MAS5 e RMA, como por exemplo o GCRMA [22], onde inclusive mostra-se que a quantidade e posição dos nucleotídeos A, T, G e C em cada sonda influencia na expressão lida pela mesma, provando a relevância da afinidade de sonda. Esta relação da afinidade da sonda em função de sua composição não é trivial: cite-se, como exemplo, a anormalmente alta afinidade de sondas com 4 guaninas consecutivas em sua composição [23]. Ao longo deste trabalho, todos os dados foram pré-processados pelo método RMA, principalmente pelo fato de este superar outros quanto à qualidade do sinal em genes com baixa expressão. Fazemos aqui uma análise de genoma completo e não desejamos realizar qualquer tipo de seleção de genes prévia à análise, como excluir aqueles com baixos níveis de expressão. Todo pré-processamento de dados foi efetuada com o pacote *affy* [24], parte integrante do projeto Biocondoctor [25], no programa estatístico R [26].

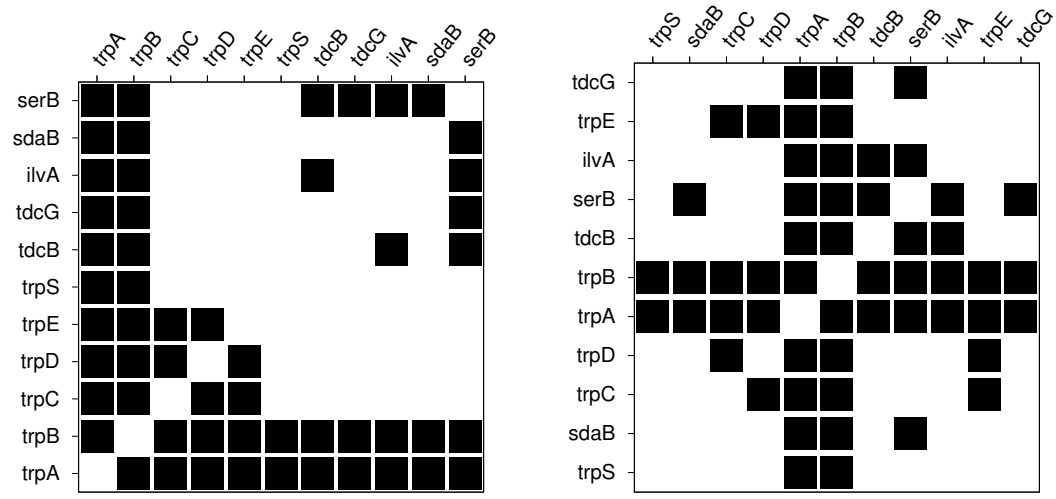
2.3 Algumas propriedades de redes

Iremos aqui descrever brevemente algumas propriedades de redes usaremos ao longo deste trabalho.

Uma rede é formada por nós associados entre si através de arestas. Vamos representar uma rede através da matriz de adjacência A, cujos termos a_{ij} são iguais a 1 se o i-ésimo nó da rede está ligado ao j-ésimo nó. Por exemplo, a rede da figura 2.1 é representada pela matriz

$$A = \begin{pmatrix} a_{N,1} & a_{N,2} & \cdots & a_{8,8} \\ \vdots & \vdots & \ddots & \vdots \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ a_{1,1} & a_{1,2} & \cdots & a_{1,8} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \qquad (2.9)$$

cujos genes estão dispostos em lista na ordem disposta na figura 2.2a. A figura 2.2



(a) (b)

Figura 2.2: Duas possíveis configurações para a matriz de adjacência para a rede mostrada na figura 2.1. Um ponto preto indica interação entre os elementos i e j da rede ($a_{ij} = 1$).

mostra a representação gráfica que usaremos para a matriz de adjacência: pontos pretos indicam interação, pontos brancos indicam ausência de interação. Nela, observam-se duas possíveis configurações para a matriz de adjacência, baseado na ordem em que estão listadas as proteínas. O método apresentado no capítulo seguinte irá estabelecer como escolher uma boa configuração para a matriz de associação, e define-se aqui que, por simplicidade, estaremos usando a palavra "ordenamento" para nos referir a uma possível configuração da matriz, isto é, a uma possível ordem de genes nesta lista.

Neste trabalho, consideramos apenas redes não-direcionadas, ou seja, simétricas: $a_{ij} = a_{ji}$. Usaremos o termo "associados" ao nos referir aos genes nas posições i e j para os quais $a_{ij} = 1$, e o termo "vizinhos" para os genes que se encontram próximos na lista. Segue agora algumas definições de rede que serão usadas na sequência deste trabalho. A conectividade k de um nó,

$$k_i = \sum_{j=1}^{N} a_{ij}, \tag{2.10}$$

mede o número de ligações em que este participa. Podemos agora calcular a distribuição de conectividades em uma rede,

$$n(k) = \sum_{i=1}^{N} \delta(k_i, k), \tag{2.11}$$

$$p(k) = \frac{n(k)}{N}, \tag{2.12}$$

além da assortatividade S, que mede qual é a conectividade média dos nós ligados a um

nó com conectividade k:

$$S(k) = \frac{\sum_{i=1}^{N} \delta(k_i,k)(\sum_{j=1}^{N} k_j a_{ij})}{n(k)k}. \tag{2.13}$$

O coeficiente de clusterização C_i, por sua vez, mede o quanto os nós associados a um nó i estão próximos de formar uma rede completa independente. É definido como a fração de ligações que os associados ao i-ésimo nó fazem entre si em relação ao total de ligações a que estes participam,

$$C_i = \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} a_{ij}a_{jk}a_{ik}}{k_i(k_i-1)}. \tag{2.14}$$

3 O ordenamento de uma rede de proteínas

Este capítulo se dedica a explicar o método de tratamento de dados de transcriptoma, o transcriptograma. Inicialmente, descrevemos o ordenamento de uma lista de proteínas com base na rede de interações entre estas para em seguida apresentar o transcriptograma. Iremos ainda caracterizá-lo a partir de critérios físicos e biológicos, apesar de isto não informar sobre o quão eficiente este ordenamento é ao que se propõe, mas se faz necessário para melhor compreender suas características.

3.1 O método

A partir do STRING, conforme descrito no capítulo anterior, obtemos todas as interações entre proteínas de determinado organismo cujo *score* de confiança seja igual ou superior a 0.8. A partir desta, monta-se uma lista, com posições aleatórias, de todas as N proteínas que participam de ao menos de uma interação, para então montarmos uma matriz de adjacência para esta lista, sendo as posições i e j da matriz correspondentes às proteínas na i-ésima e j-ésima posições na lista. Definimos então uma energia para esta matriz que será tanto maior quanto maior for a presença de interfaces entre 1's e 0's e quanto maiores forem as distâncias destas interfaces até a diagonal da matriz, isto é,

$$E = \sum_{i,j} a_{ij} D_{i,j} I_{i,j}, \tag{3.1}$$

onde $D_{i,j}$ é o termo de distância da posição do elemento a_{ij} até a diagonal da matriz, dado por

$$D_{i,j} = |i - j|, \tag{3.2}$$

e $I_{i,j}$ é o termo de interfaces,

$$I_{i,j} = |a_{i,j} - a_{i,j-1}| + |a_{i,j} - a_{i,j+1}| + |a_{i,j} - a_{i-1,j}| + |a_{i,j} - a_{i+1,j}|. \tag{3.3}$$

O termo a_{ij} na equação 3.1 cumpre a função de impedir que uma interface seja con-

tada duas vezes. De fato, uma vez que a_{ij} assume apenas os valores 1 e 0, sua presença apenas indica que D_{ij} e I_{ij} serão apenas calculados caso as proteínas localizadas nas posições i e j interajam. Isto representa uma grande economia no tempo de execução do algoritmo de ordenamento.

Em seguida, a lista de proteínas é ordenada via Monte Carlo com *annealing* simulado [27], como segue. Duas proteínas aleatoriamente escolhidas têm suas posições na lista invertidas, dando origem a uma nova matriz M e a uma nova energia E. Esta nova configuração do sistema é aceita caso a variação de energia δE entre a nova e antiga configuração satisfizer $\delta E < 0$. No caso em que $\delta E \geqslant 0$, a nova configuração é aceita com probabilidade $\exp -\delta E/T$, onde T é uma temperatura virtual. Para este trabalho, todos os ordenamentos foram produzidos com a temperatura inicial definida como 1% da energia total do sistema e reduzida a 50% de seu valor anterior a cada 200 passos de Monte Carlo. Um passo de Monte Carlo corresponde a N tentativas de troca, e o processo é repetido por 100 reduções de temperatura. Neste trabalho, vamos generalizar a energia apresentada na equação 3.1 de modo a balancear a importância que damos para o termo de interfaces e o de distância com

$$E = \sum_{i,j} a_{ij} D_{ij}^{\alpha} I_{ij}^{\beta}. \tag{3.4}$$

Este método para ordenamento de redes foi previamente apresentado [28] [1], sendo também generalizado para duas dimensões [29].

Apresentamos gráficos das matrizes de adjacências ordenadas para a rede do *Homosapiens* nas figuras 3.1, 3.2, 3.3 e 3.4, onde observa-se a influência dos parâmetros α e β na estrutura final destas. Observamos que com um valor alto de α obtemos um formato de folha, por assim dizer, onde proteínas que interagem são impedidas de ficar muito distantes no ordenamento, visto a total ausência de pontos pretos longe da diagonal da matriz. Em contrapartida, para os parâmetros $\alpha = 1$ e $\beta = 1$, prioriza-se a formação de grupos de genes que apresentam forte conexão interna, visto a presença de grupos de pontos pretos próximos à diagonal.

3.1.1 Transcriptograma

Enfim, a técnica de análise de expressão gênica, o transcriptograma, alvo do presente estudo, é aqui apresentado. Atribuímos, inicialmente, a cada gene da lista o seu valor de expressão, independente para cada amostra realizada. O método consiste em realizar médias sobre o valor de expressão de genes vizinhos no ordenamento. Baseado na hipótese de que um ruído, de caráter aditivo e descorrelacionado do sinal, está pre-

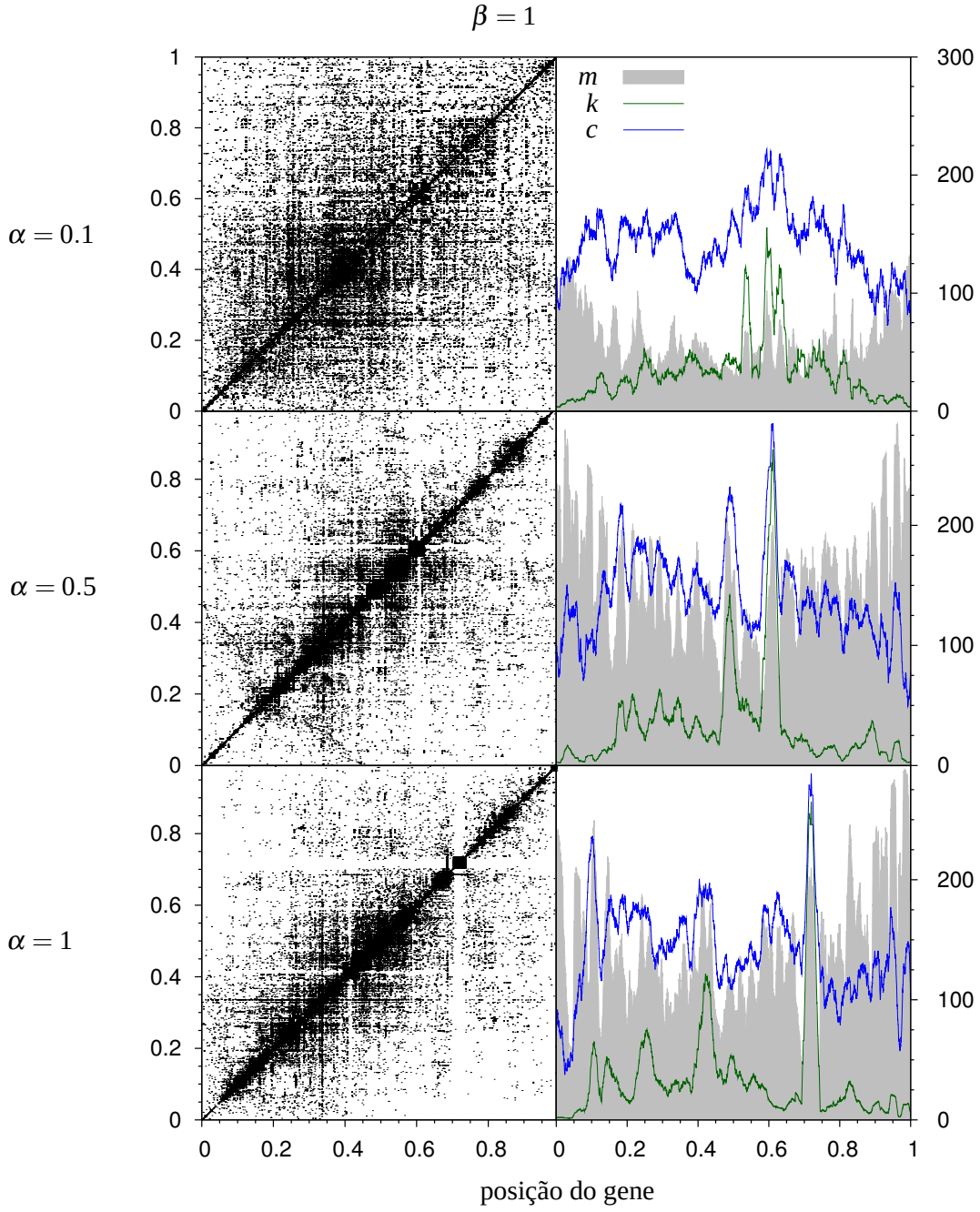


Figura 3.1: Ordenamentos da rede do *Homo sapiens*, mostrando matrizes de adjacência à esquerda e perfis de modularidade m, conectividade k e clusterização C. Eixo vertical da esquerda: posição (matriz), modularidade e clusterização. Eixo vertical da direita: conectividade.

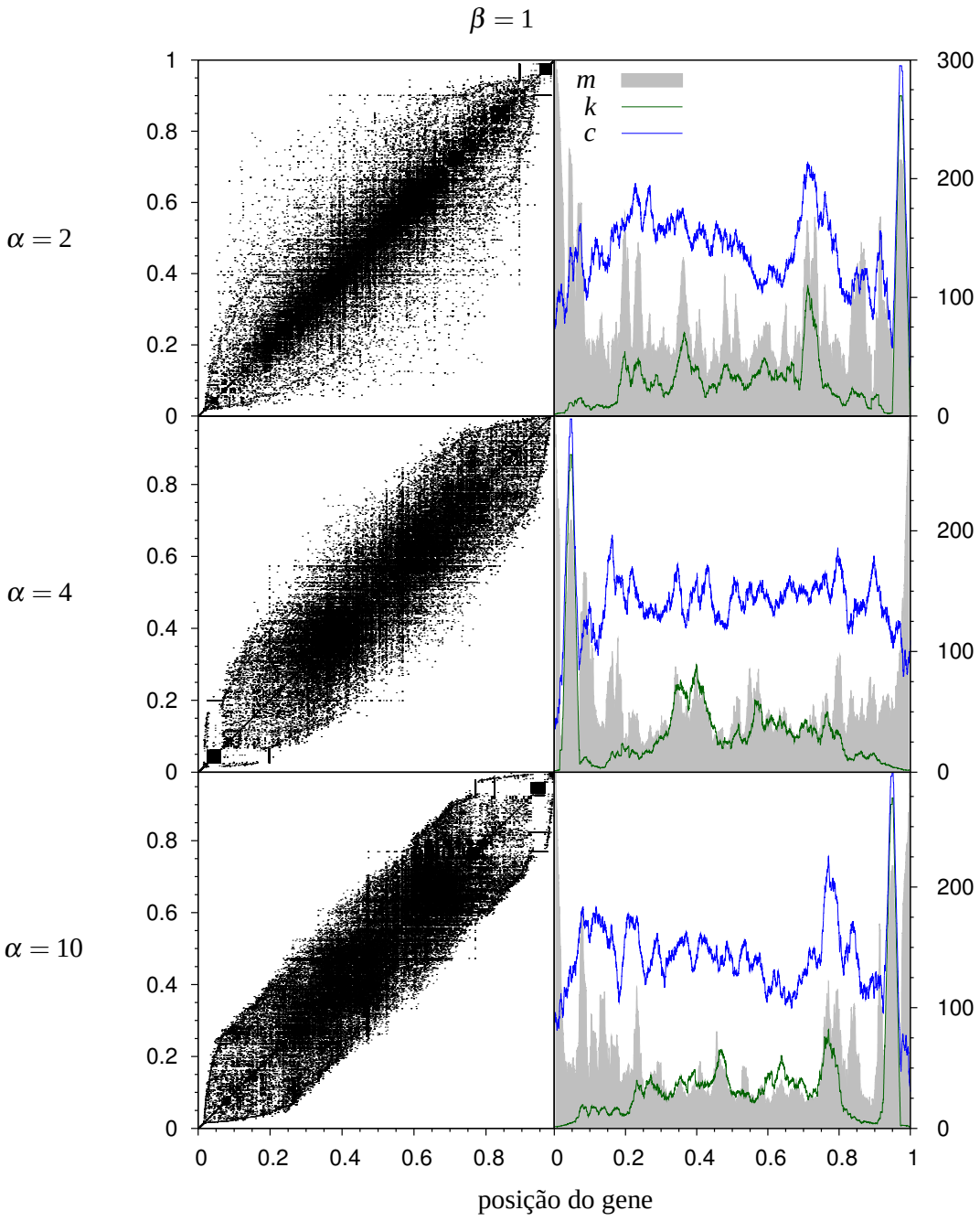


Figura 3.2: Ordenamentos da rede do *Homo sapiens*, mostrando matrizes de adjacência à esquerda e perfis de modularidade m, conectividade k e clusterização C. Eixo vertical da esquerda: posição (matriz), modularidade e clusterização. Eixo vertical da direita: conectividade.

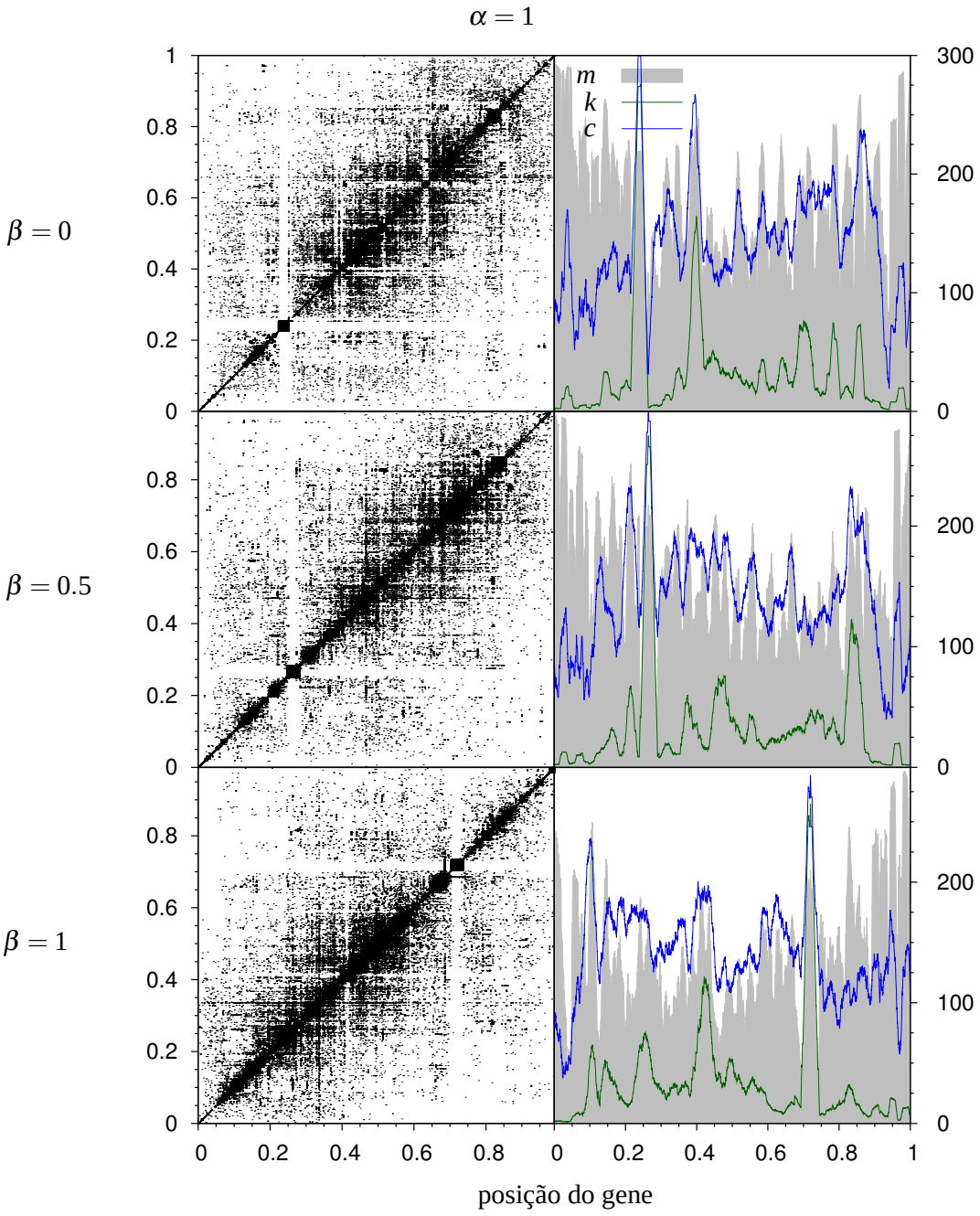


Figura 3.3: Ordenamentos da rede do *Homo sapiens*, mostrando matrizes de adjacência à esquerda e perfis de modularidade m, conectividade k e clusterização C. Eixo vertical da esquerda: posição (matriz), modularidade e clusterização. Eixo vertical da direita: conectividade.

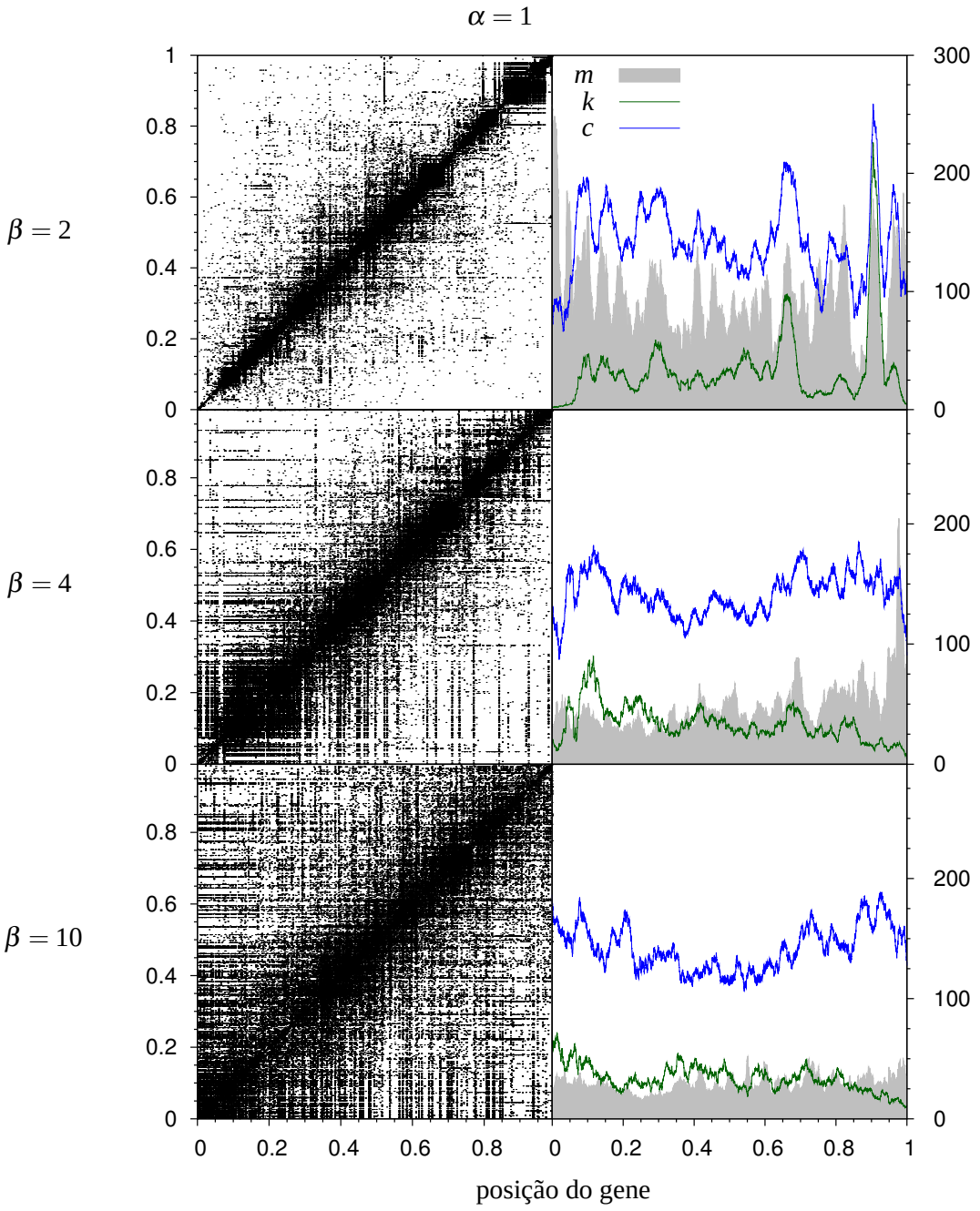


Figura 3.4: Ordenamentos da rede do *Homo sapiens*, mostrando matrizes de adjacência à esquerda e perfis de modularidade m, conectividade k e clusterização C. Eixo vertical da esquerda: posição (matriz), modularidade e clusterização. Eixo vertical da direita: conectividade.

sente a medida, a média da expressão sobre uma vizinhança que apresenta correlação do seu sinal produz um perfil de expressão onde minimizamos a intensidade do ruído, preservando o sinal daquele grupo de genes. Observe o leitor que, nos métodos de pré-processamento de dados de microarranjo MAS5 e RMA, descritos no capítulo anterior, supõe-se que o ruído da medida é multiplicativo ao sinal, e que o RMA produz um valor de expressão em escala logarítmica. Como dito anteriormente, todo transcriptoma analisado neste trabalho é pré-processado através deste último e, portanto, estamos por hipótese tratando de um erro aditivo nesta escala. Definimos, assim, o valor do transcriptograma τ para o gene i como

$$\tau_i = \langle t \rangle_{w_i} = \frac{\sum_j H(r - d_{ij}) t_j}{\sum_j H(r - d_{ij})}, \tag{3.5}$$

onde t_i é o valor de transcrição, ou expressão, do gene situado na posição i, w_i denota a janela ou região de raio r sobre a qual realizamos a média de expressão, e termo d_{ij} indica a distância entre o i-ésimo e o j-ésimo genes da lista e

$$H(x) = \begin{cases} 0 \text{ se } x < 0 \\ 1 \text{ se } x \geq 0 \end{cases} . \tag{3.6}$$

Por si só, o melhor valor possível de r, ou seja, o tamanho da janela ou vizinhança que estamos usando para suavizar o perfil de expressão é uma questão em aberto.

Apesar de este método já ter sido utilizado, ainda não há uma análise explícita sobre qual o seu comportamento sobre a redução do ruído de medida. Ainda, toda análise até então efetuada com o transcriptograma foi apenas para um ordenamento produzido com os parâmetros $\alpha = 1$ e $\beta = 1$. E uma pergunta relevante ainda se mantém, que é a de quão adequado é esta escolha desta vizinhança sobre a qual estamos realizando médias. Atente-se ao importante fato de que um "bom transcriptograma" é uma termo de difícil definição, visto que a real concentração dos RNA's presentes na amostra nos é desconhecido. Não podemos comparar diretamente o transcriptograma obtido com algum tipo de padrão de ouro. Tais questões ainda serão abordadas em capítulos subsequentes, mas antes ainda precisamos caracterizar os ordenamentos que produzimos, tema da próxima seção. Todas os transcriptogramas que seguem neste trabalho foram feitos para *Homo sapiens*.

3.2 Caracterização

3.2.1 Modularidade, conectividade e clusterização

Um módulo funcional é um grupo de genes altamente interligados e que se compartilham características biológicas [30], participando das mesmas ontologias. Para caracterizar o quão interagente estão os genes localizados próximos na lista ordenada, definimos a medida de modularidade da janela, m_i, para todo gene i,

$$m_i(r) = \frac{\sum_{j=1}^{N} H(r-d_{ij}) \sum_{k=1}^{N} H(r-d_{ik}) a_{jk}}{\sum_{j=1}^{N} H(r-d_{ij}) \sum_{k=1}^{N} a_{jk}}, \tag{3.7}$$

que mede o quanto os genes presentes em uma região centrada em i e composta por genes que distam até r posições na lista estão interconectados entre si em relação ao total de conexões de que participam.

Ainda, uma vez que podemos calcular a conectividade e o coeficiente de clusterização de cada gene da rede, podemos produzir um perfil destes ao longo do ordenamento. Produzimos este perfil como se realizando um transcriptograma, ou seja, realizando médias sobre o valor da conectividade sobre genes em uma janela de raio r. Vemos nas figuras 3.1, 3.2, 3.3, 3.4, à direita, a modularidade e os perfis de conectividade e clusterização para os ordenamentos realizados. Observa-se que regiões onde há presença de pontos pretos próximos à diagonal é marcado por um pico de modularidade, geralmente acompanhado de picos dos perfis de conectividade e clusterização da rede. Note que regiões de muito baixa conectividade correspondem a alta modularidade, o que ocorre devido ao fator de normalização no denominador da equação 3.7, o que geralmente é o caso para os extremos do ordenamento. É uma consequência do método, que não usa condições de contorno para o cálculo da distância, que genes de baixa conectividade acabem se dirigindo nos extremos da lista.

3.2.2 Nível de ocupação

De modo a analisar como os genes estão próximos daqueles aos quais estão associados, podemos definir, para cada proteína i, o nível de ocupação dos vizinhos do i-ésimo gene,

$$o_i(d) = \frac{1}{k_i} \sum_{j=1}^{N} \delta(d_{ij}, d) a_{ij}, \tag{3.8}$$

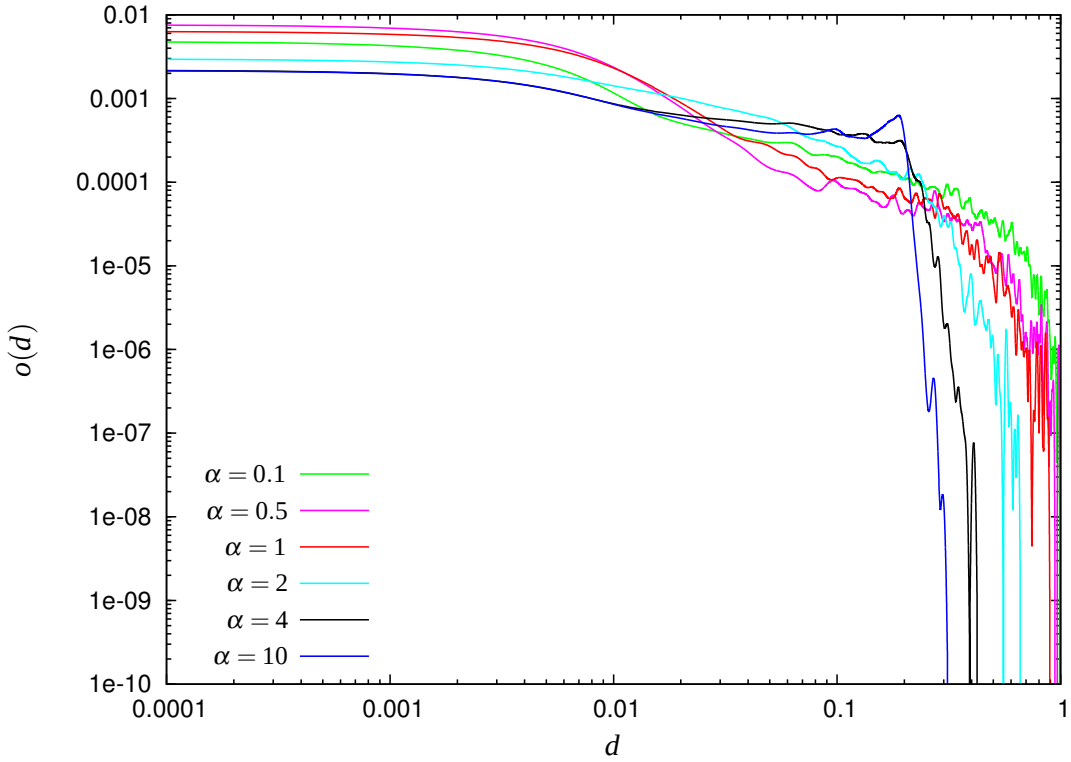


Figura 3.5: Níveis de ocupação para $\beta = 1$, com distância normalizada pelo tamanho da rede.

onde entenda-se por nível de ocupação a fração, do total de genes que estão a uma distância d, daqueles com quem este está associado. Agora podemos calcular a ocupação média para todos os genes,

$$o(d) = \frac{1}{N}\sum_{i=1}^{N} o_i(d). \tag{3.9}$$

Nas figuras 3.5 e 3.6, mostramos os perfis de ocupação para os ordenamentos previamente mostrados. Nota-se a marcação da estrutura de folha para $\alpha = 10$ inclusive havendo um aumento da ocupação antes da queda abrupta e que maiores valores de α causa a queda da ocupação a zero a distâncias cada vez menores. Esta queda não acontece para o caso $\alpha = 1$ e $\beta = 1$, por exemplo, havendo associação em distâncias até próximo ao tamanho da rede. Observa-se porém que o formato de folha acarreta em uma menor ocupação para distâncias menores, mostrando que uma escolha de vizinhança para a realização do transcriptograma corresponde, neste caso, a escolha de genes menos associados entre si.

3.2.3 Caracterização biológica

Vamos agora realizar a caracterização biológica do ordenamento. Com base em dados obtidos a partir do G*ene Ontology*, podemos obter listas, para cada processo biológico, dos genes que deste participam. Marcamos então, no ordenamento, com 1 todo

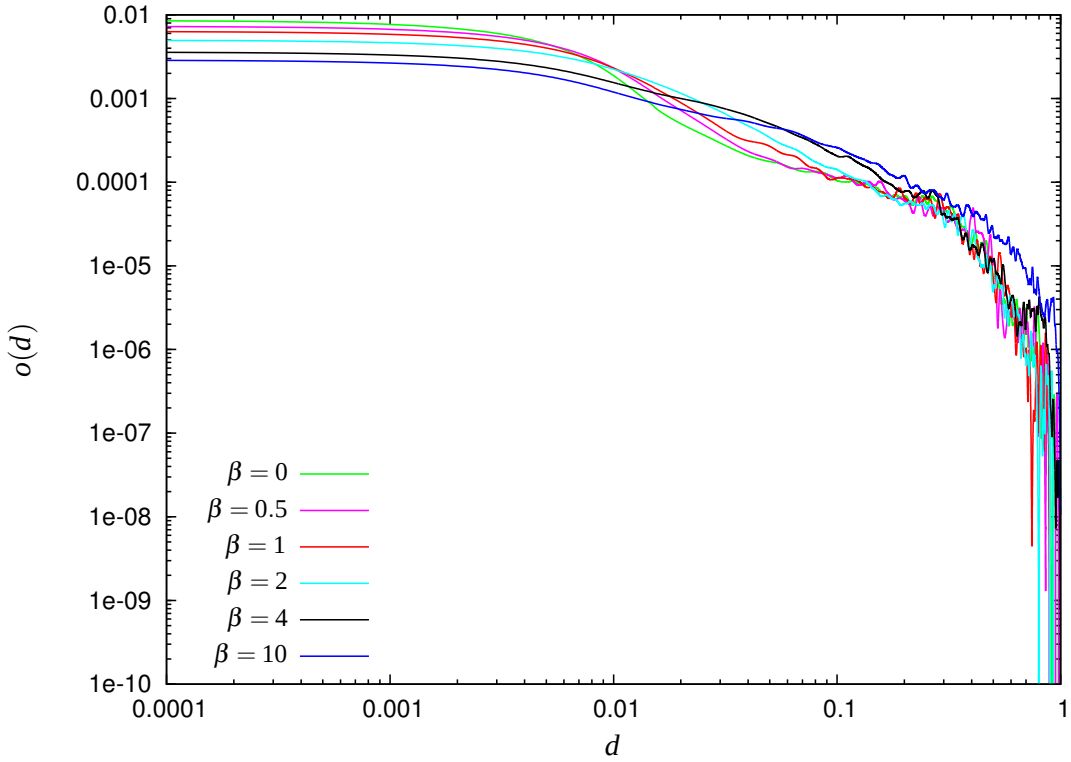


Figura 3.6: Níveis de ocupação para $\alpha = 1$, com distância normalizada pelo tamanho da rede.

gene que participe de cada processo biológico e com 0 para aqueles que não participam, para em seguida efetuar um transcriptograma do processo biológico, por assim dizer, realizando médias sobre um janela de raio r. Em outras palavras, estamos calculando uma densidade de participação dos genes em cada processo biológico, ao qual chamaremos de enriquecimento funcional. Tais perfis de densidade podem ser vistas nas figuras 3.7, 3.8, 3.9 para $\alpha = 1$ e $\beta = 1$, onde observa-se que genes que participam das mesmas funções biológicas tendem a ficar próximos. Nestas figuras, a modularidade também é mostrada para guiar o olho. De fato, tal correlação da participação em processos biológicos com a distância justifica o ordenamento da rede de interação proteica a fim de gerar um transcriptograma. Dado que a lista de genes agora aproxima aqueles que mais provavelmente participam da mesma função biológica, podemos esperar uma certa correlação na sua atividade transcricional.

Precisamos agora comparar como se dá esta projeção de processos biológicos para distintos ordenamentos. Para isso, calculamos a curtose [31], que é uma medida do quanto uma distribuição está concentrada ao redor da média. É definido como

$$\gamma = \frac{\mu_4}{\sigma^4}, \tag{3.10}$$

onde μ é o quarto momento central da distribuição com N elementos x_i, e σ é o desvio

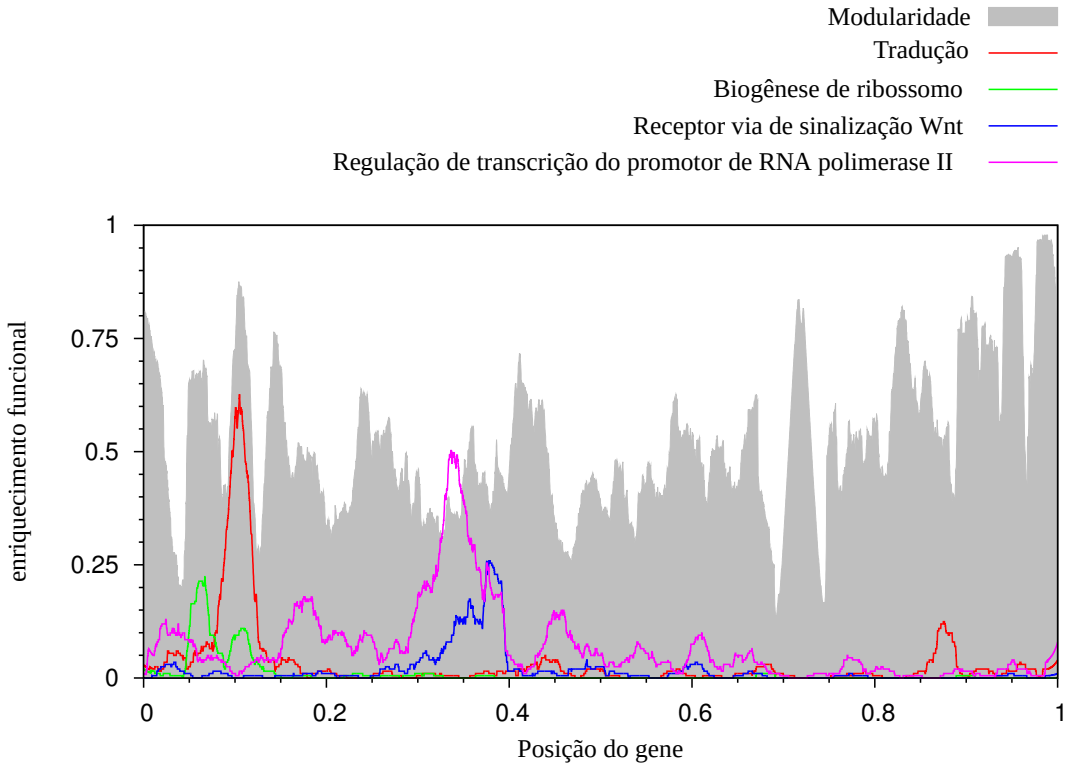


Figura 3.7: Alguns exemplos de ontologias, localizados à esquerda do ordenamento para $\alpha = 1$ e $\beta = 1$.

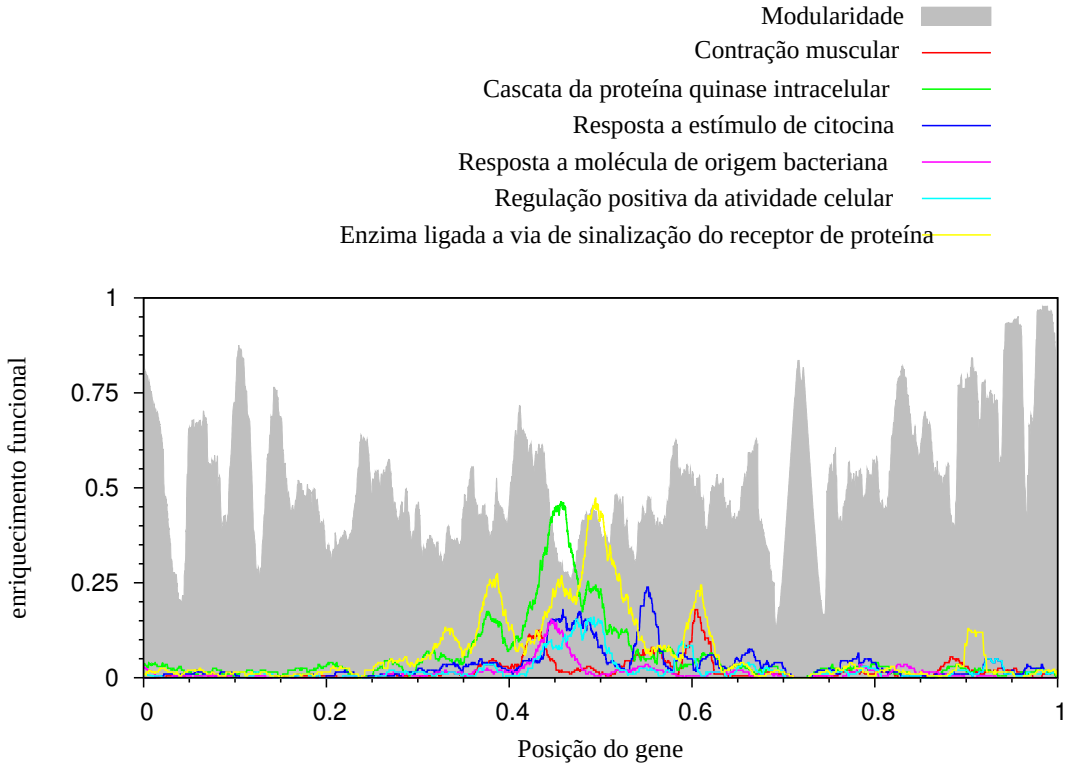


Figura 3.8: Alguns exemplos de ontologias, localizados mo centro do ordenamento para $\alpha = 1$ e $\beta = 1$.

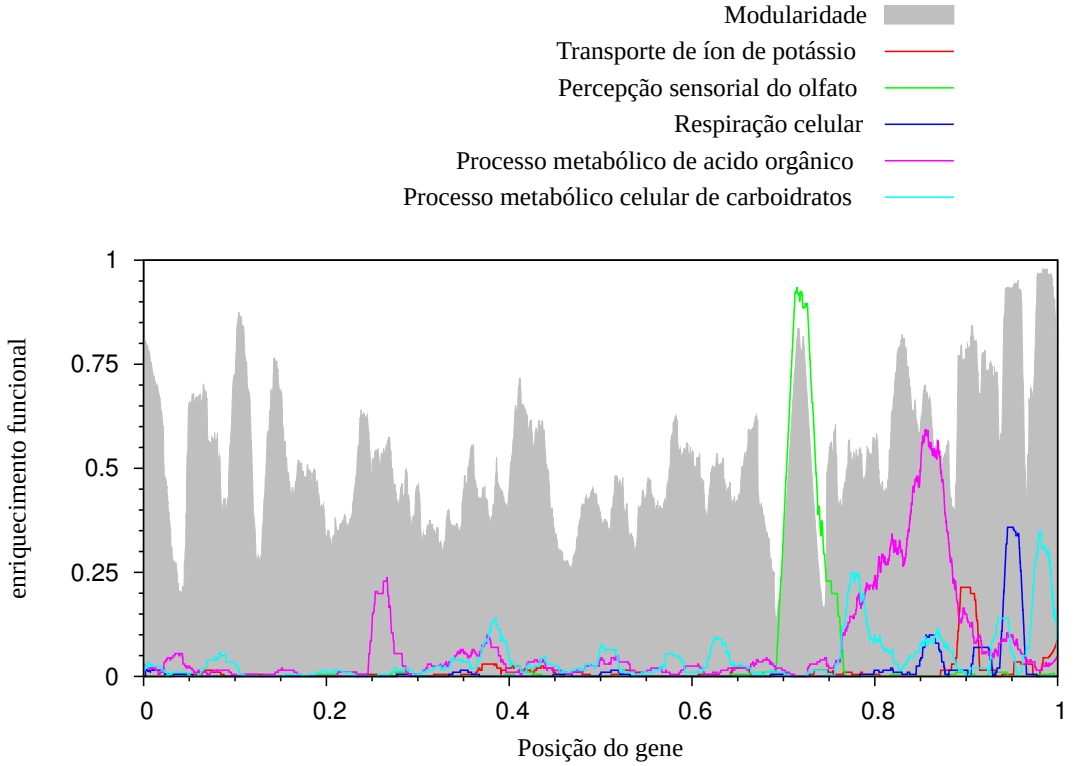


Figura 3.9: Alguns exemplos de ontologias, localizados à direita do ordenamento para $\alpha = 1$ e $\beta = 1$.

padrão,

$$\mu_4 = \sum_i \frac{(x_i - \bar{x})^4}{N}\ ,\ \sigma = \sqrt[2]{\sum_i \frac{(x_i - \bar{x})^2}{N}}. \tag{3.11}$$

Para uma distribuição normal, $c = 3$, enquanto que para uma distribuição uniforme, $c = 1.8^*$, sendo que quanto maior a curtose, mais intensa é a concentração dos seus elementos em torno da média. Podemos então calcular a curtose c para cada processo biológico para podermos comparar como este se comporta em diferentes ordenamentos, bem como a altura máxima de cada processo biológico, que chamaremos de P_m. Para fazer isso em larga escala, fazemos a média destes parâmetros, $\overline{\gamma}$ e $\overline{P_m}$, para 146 processos biológicos, listados no apêndice B. Observamos nas figuras 3.10 e 3.11 que um ordenamento com $\alpha = 1$ e $\beta = 0$ possui uma maior capacidade de aglutinar genes com a mesma função biológica.

O transcriptograma, descrito neste capítulo, é concebido a partir da justificativa de que aproxima genes com funções semelhantes bem como a de correlacionar distância com probabilidade de associação, o que justificaria o ordenamento como a adequada escolha de vizinhanças sobre as quais se deve realizar médias sobre sua expressão. Mostra-se aqui, no entanto, que a ausência do termo de interfaces produz um ordena-

*Comumente prefere-se uma definição alternativa de curtose, $c = \frac{\mu_4}{\sigma^4} - 3$, assim feita de modo que a curtose de uma curva normal seja 0. Como isto não afeta o que é demonstrado neste trabalho, usamos a definição apresentada no texto.

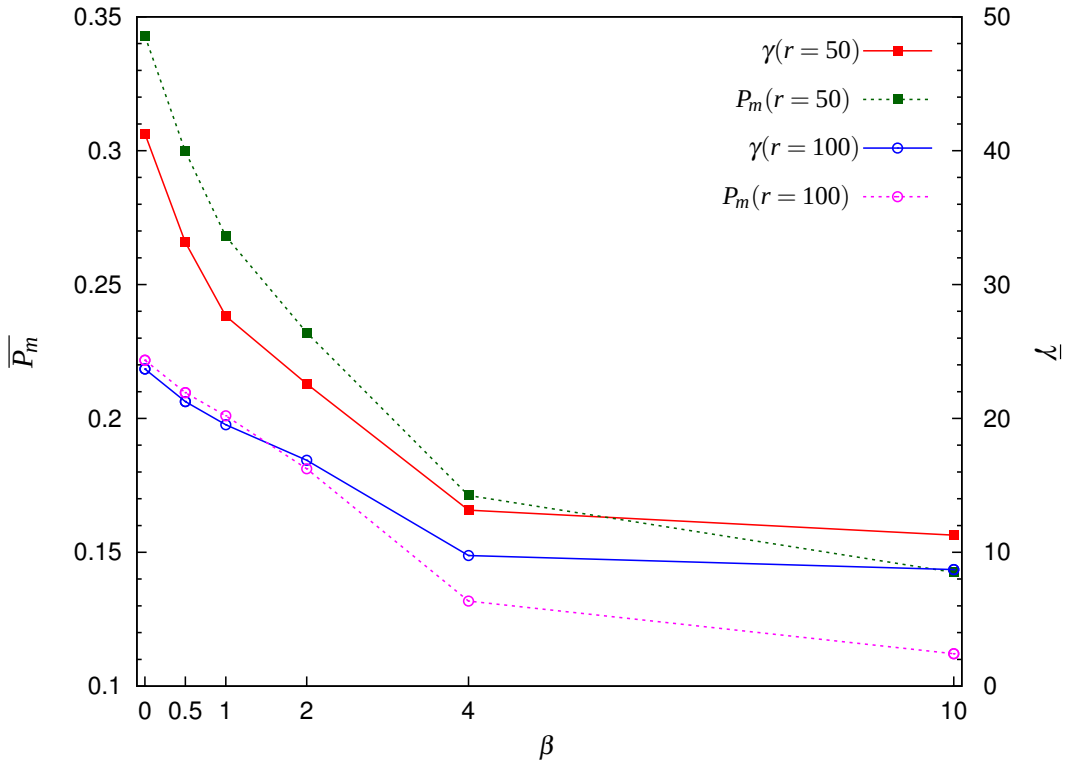


Figura 3.10: Altura máxima e curtose para ordenamentos como função de β com $\alpha = 1$. O eixo vertical da direita corresponde a P_m, enquanto que o da esquerda, a γ

mento que correlaciona ainda mais proximidade na lista com função biológica. Ainda, observamos que o ordenamento com a chamada estrutura de folha destrói substancialmente tal correlação, apesar de tal ordenamento impedir que pares associados distam mais do que uma certa distância limite. Até o momento, apenas justificamos o transcriptograma baseado em hipóteses, mas nada foi dito se estamos de fato obtendo algum tipo de sucesso no que inicialmente propomos, o de analisar o sinal de um transcriptoma.

O ordenamento de redes pode servir a mais de um fim, e se este for o de analisar o sinal de expressão de uma amostra puramente com base nas funções biológicas dos genes, está aqui demonstrado que o conjunto de parâmetros até então utilizado em todas as análises por transcriptograma, $\alpha = 1$ e $\beta = 1$, não é aquele de melhor resultado. A exclusão do termo de interface na energia, equação 3.1, é portanto mais indicado para o caso. No entanto, é a partir do sinal do transcriptograma que vamos basear o estudo que se verá mais adiante. Estamos aqui analisando o quão eficaz é um transcriptograma ao produzir um perfil de expressão que melhor caracterize uma amostra sem nos basear *a priori* em nenhuma característica biológica de determinado ordenamento.

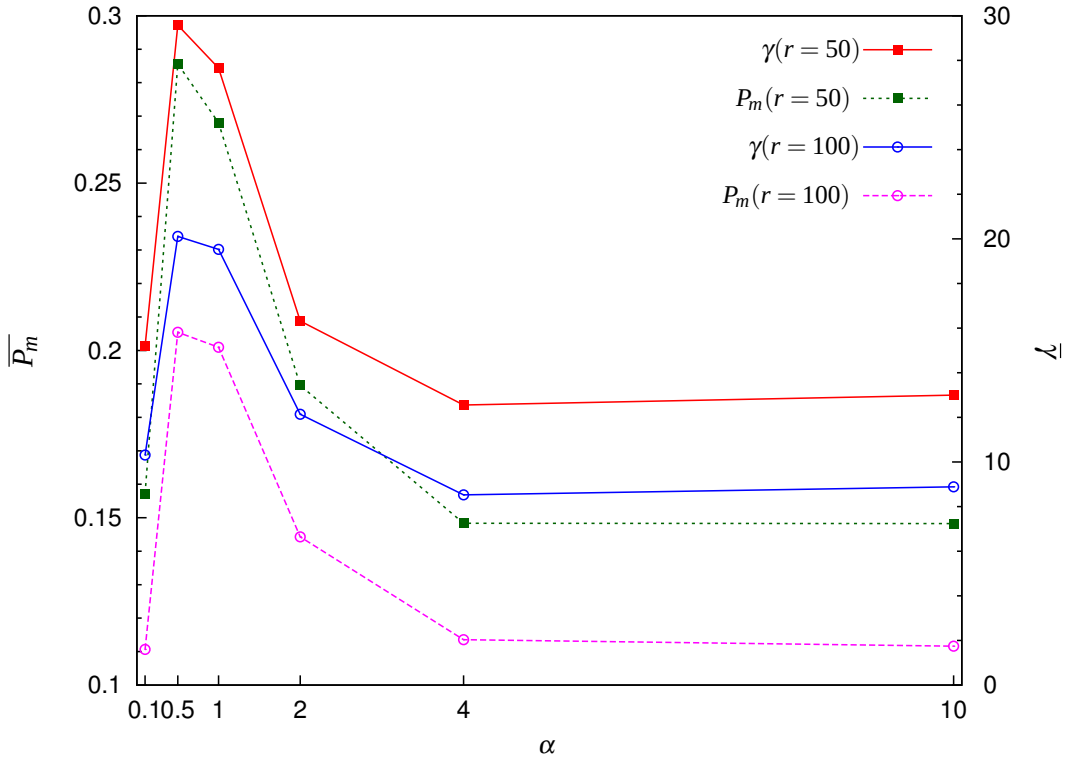


Figura 3.11: Altura máxima e curtose para ordenamentos como função de α com $\beta = 1$. O eixo vertical da direita corresponde a P_m, enquanto que o da esquerda, a γ

4 *Sobre o transcriptograma e o ruído da medida*

Apesar de o transcriptograma já ter se mostrado útil em pesquisa anterior, quando da investigação do ciclo celular da *Saccharomyces cerevisiae* [1], inexiste até o presente momento uma comprovação de que o transcriptograma é efetivo em retirar ou reduzir uma variação irreal presente na medida, que não corresponde à verdadeira expressão dos genes do organismo. Quanto maior for o raio r de suavização do sinal em um transcriptograma τ, equação 3.5, é evidente que mais suave será o perfil de expressão ao longo do ordenamento, como mostrado na figura 4.1. A atenuação do sinal é inevitável, e para raios muito grandes começamos a perder a correlação na expressão dos genes dentro da mesma janela. De fato, se colocarmos os genes em posições aleatórias no ordenamento, um transcriptograma sobre tal ordenamento produz um sinal que tende a média da expressão global muito mais rápido (figura 4.1, curva em azul). O quanto estamos perdendo informação em relação a qualidade do sinal que obtemos é alvo do estudo no capítulo seguinte. Neste capítulo, nos centraremos na questão de se um transcriptograma é realmente efetivo em reduzir o ruído da medida do transcriptoma.

4.1 Ruído

De modo a analisar a presença e comportamento do ruído de uma medida ao se realizar um transcriptograma, analisamos aqui o experimento registrado sobre o código GSE26244[32] no GEO, o qual apresenta 6 replicatas técnicas de umas das amostras. Uma replicata técnica é aquela em que uma mesma amostra de RNA é hibridizada em mais de um microarranjo. Portanto, toda diferença na medida de expressão entre replicatas técnicas se origina tão somente a partir na técnica da medida. Usualmente, na literatura, se realizam replicatas biológicas das amostras analisadas, onde todo o tratamento de produção da amostra em si é feito mais de uma vez, com o objetivo de analisar também a variação biológica. O experimento GSE26244 foi escolhido tão somente devido ao incomum alto número (6) de replicatas técnicas, mais adequado à análise descrita

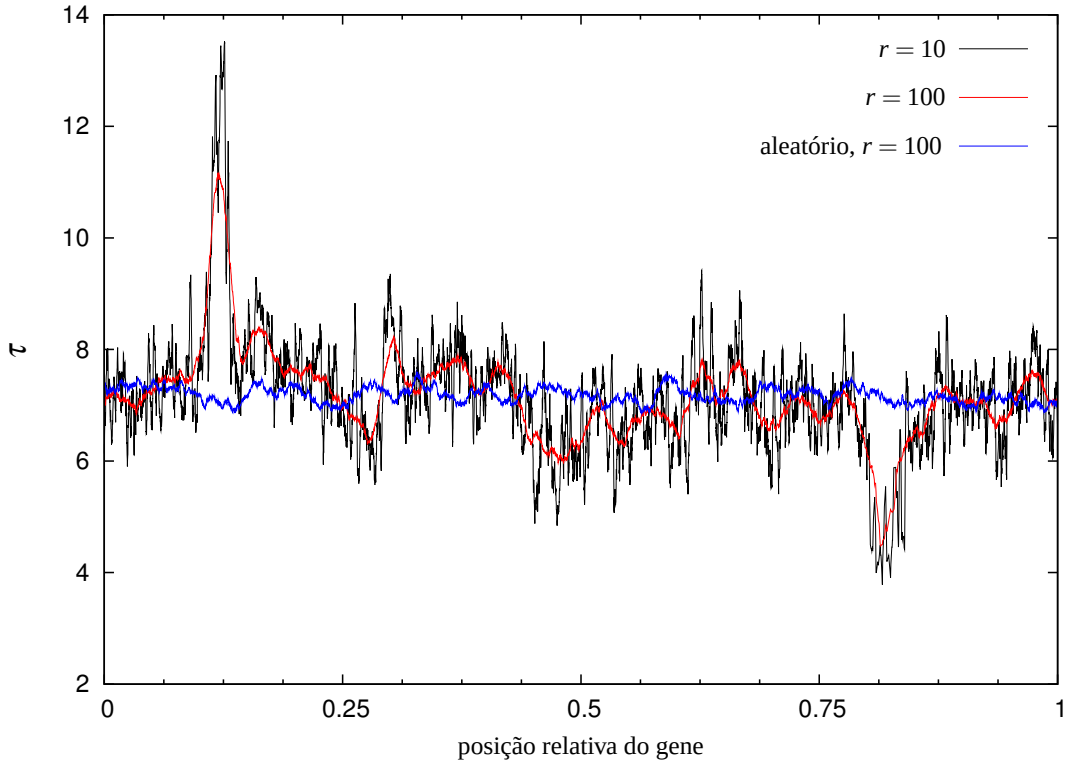


Figura 4.1: Transcriptograma τ para uma amostra de microarranjo do experimento cadastrado no GEO pelo código GSE13355 [5] [6], para dois raios diferentes em um ordenamento $\alpha = 1$, $\beta = 1$, bem como para um ordenamento aleatório (azul).

a seguir.

O que propomos é a análise da variação de um transcriptograma sobre amostras que deveriam ser idênticas. Por hipótese, o sinal medido pelo transcriptoma, que lembramos está em escala logarítmica pois os dados foram pré-processados por RMA, para o gene i, t_i, pode ser descrito como a soma do sinal real, s_i, e um ruído de medida, intrínseco à técnica, ξ_i, de caráter aditivo:

$$t_i = s_i + \xi_i, \tag{4.1}$$

sendo ξ_i e s_i descorrelacionados, com ξ_i possuindo uma distribuição normal. O transcriptograma é obtido realizando a média de t_i em uma janela ω, como já definido na eq. 3.5,

$$\tau_i = \widetilde{t_i} = \widetilde{s_i} + \widetilde{\xi_i}. \tag{4.2}$$

Calculamos agora médias e desvios de transcriptogramas sobre replicatas técnicas,

$$\langle \tau_i \rangle_{rep} = \langle \widetilde{s_i} \rangle_{rep} + \langle \widetilde{\xi_i} \rangle_{rep}, \tag{4.3}$$

$$\langle \tau_i \rangle^2_{rep} = \langle \widetilde{s_i} \rangle^2_{rep} + \langle \widetilde{\xi_i} \rangle^2_{rep} + 2\langle \widetilde{s_i} \rangle_{rep} \langle \widetilde{\xi_i} \rangle_{rep}, \tag{4.4}$$

$$\tau_i^2 = \widetilde{s_i}^2 + \widetilde{\xi_i}^2 + 2\widetilde{s_i}\widetilde{\xi_i}, \tag{4.5}$$

$$\langle \tau_i^2 \rangle_{rep} = \langle \widetilde{s_i}^2 \rangle_{rep} + \langle \widetilde{\xi_i}^2 \rangle_{rep} + 2\langle \widetilde{s_i}\widetilde{\xi_i} \rangle_{rep}, \tag{4.6}$$

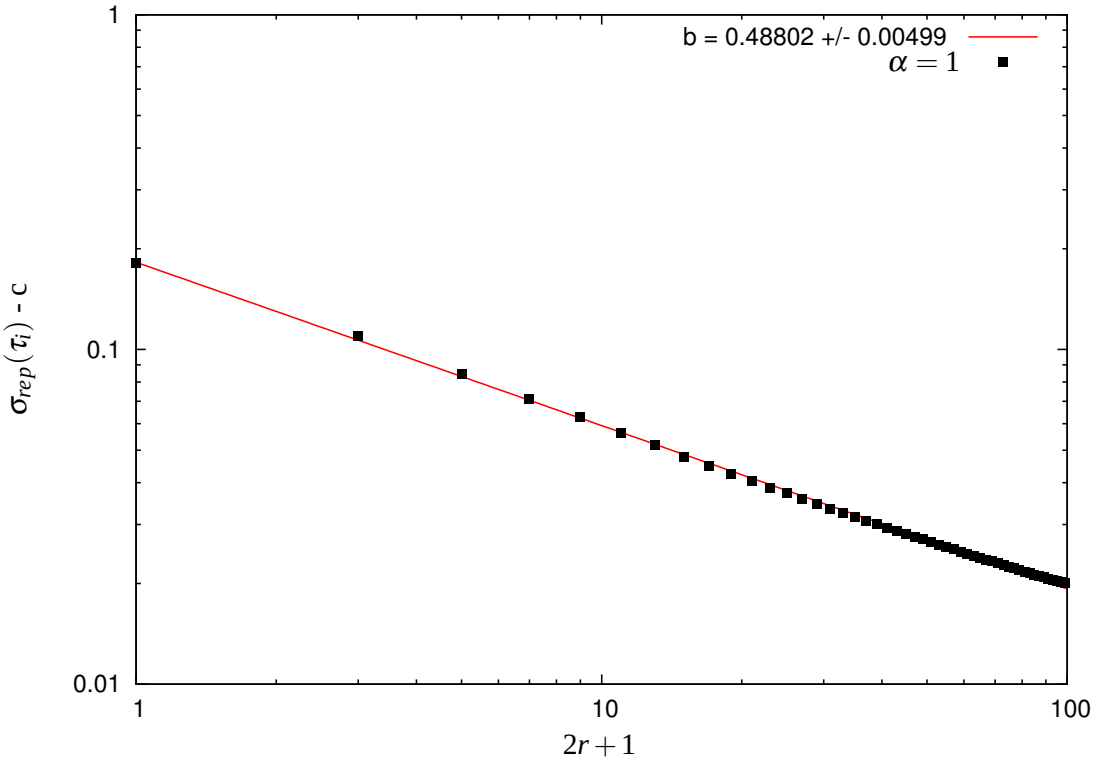


Figura 4.2: Ruído medido entre replicadas técnicas como função do tamanho da janela, subtraído da constante c obtida por ajuste da equação 4.12. Linha sólida é a curva $f(x) = a * x^b$

onde

$$< x >_{rep} = \frac{1}{n_{rep}} \sum_{j=1}^{n_{rep}} x_j, \tag{4.7}$$

com n_{rep} sendo o número de replicatas.

Usando que s_i e $\widetilde{s_i}$ são, por hipótese, invariantes sobre replicatas e que s_i e ξ_i são descorrelacionados:

$$\langle \tau_i \rangle^2_{rep} = \widetilde{s_i}^2 + \langle \widetilde{\xi_i} \rangle^2_{rep} + 2\widetilde{s_i} \langle \widetilde{\xi_i} \rangle_{rep}, \tag{4.8}$$

$$\langle \tau_i^2 \rangle_{rep} = \widetilde{s_i}^2 + \langle \widetilde{\xi_i}^2 \rangle_{rep} + 2\widetilde{s_i} \langle \widetilde{\xi_i} \rangle_{rep}, \tag{4.9}$$

$$\sigma^2_{rep}(\tau_i) = \langle \tau_i^2 \rangle_{rep} - \langle \tau_i \rangle^2_{rep} = \langle \widetilde{\xi_i}^2 \rangle_{rep} - \langle \widetilde{\xi_i} \rangle^2_{rep}, \tag{4.10}$$

$$\sigma^2_{rep}(\tau_i) = \sigma^2_R(\widetilde{\xi_i}). \tag{4.11}$$

Ou seja, o desvio padrão entre transcriptogramas de replicatas técnicas é igual ao desvio do ruído. A partir dos transcriptogramas das 6 replicatas técnicas do experimento GSE26244, calculamos o desvio padrão para cada gene do ordenamento. A figura 4.2 mostra a relação do desvio padrão médio de todos os genes com o tamanho da janela usada para a realização do transcriptograma. Para analisar como este ruído decai com o

tamanho da janela, ajustamos a curva

$$f(x) = c + ax^b, \tag{4.12}$$

o que resulta em um expoente aproximadamente $b = 1/2$, indicando que existe uma componente no sinal que é aditivo e branco. Observa-se ainda que $c = 0.2$, indicando que existe uma diferença entre amostras que o processo de normalização do pré-processamento dos dados por RMA foi incapaz de retirar. Note, a etapa de normalização no RMA não é a última, ao contrário do que acontece no MAS5.

4.2 Reprodutibilidade

Seguindo nosso estudo sobre as vantagens do transcriptograma, vamos agora nos deter sobre uma crítica feita sobre medidas de expressão por microarranjo, que é a reprodutibilidade dos resultados. Uma medida científica deve, necessariamente, ser reproduzida em qualquer laboratório independentemente do original, obtendo-se resultados coerentes entre ambas as medidas. Devido ao ruído intrínseco à medida do microarranjo, existem críticas sobre os métodos utilizados para se obter resultados coerentes entre laboratórios. Em especial, o projeto MicroArray Quality Control (MAQC) [33] se dedicou, entre outros assuntos, a estudar tal fenômeno.

Usualmente, a análise de dados de microarranjo se dá a partir da comparação dos níveis de expressão dos genes entre dois grupos de amostras distintas, por exemplo, amostras de células tratadas com uma substância química contra amostras controle (não tratadas), células cancerosas contra células normais, etc. Esta comparação normalmente consiste na expressão relativa ER, que nada mais é do que a razão da expressão média de cada gene entre os dois grupo, ou o p-valor de um teste-t de Student [34]. Seja a hipótese de que a média da expressão de um gene é igual nos dois grupos de amostras, o p-valor de um teste-t de Student é a probabilidade de se obter uma observação tão ou mais extrema do que as populações de fato observadas. Em outras palavras, quanto menor é o p-valor de determinado gene, menos provável é a hipótese de que a sua expressão não variou entre as duas condições em análise.

O quão bem estes dois métodos distintos, de identificação de genes diferencialmente expressos em duas condições distintas, produzem informação confiável é alvo de estudo em, por exenplo, [35] e [36], e é este o alvo da discussão a que se dedica esta seção. Em um experimento cadastrado no GEO pelo código GSE41328 [37], dois laboratórios diferentes realizaram, para as mesmas amostras, medidas de transcriptoma por microarranjo. As amostras em questão são de adenocarcinoma colorretal e de tecido saudável,

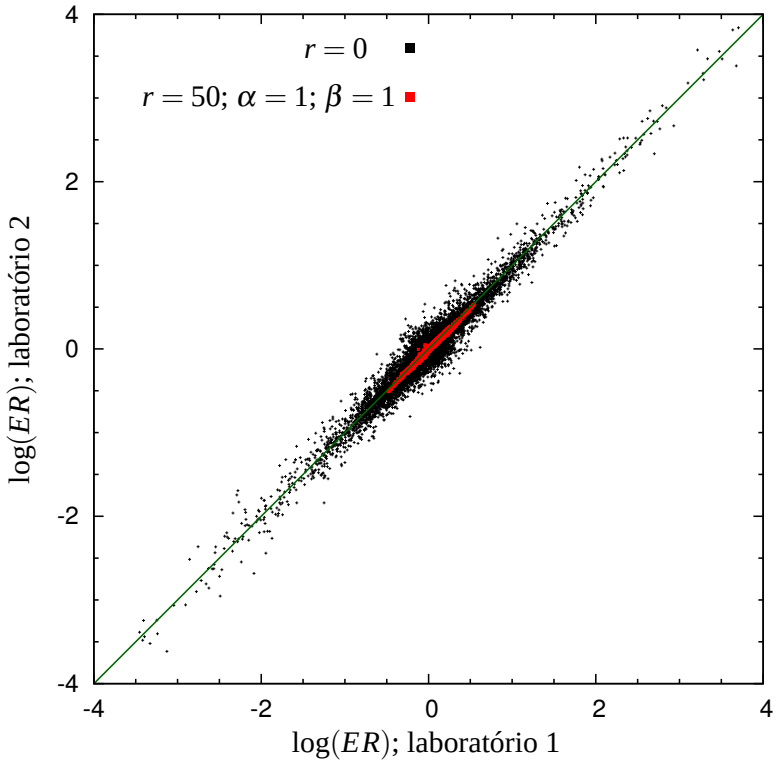


Figura 4.3: Reprodutibilidade entre laboratórios da expressão relativa ER para transcriptoma (preto) e transcriptograma (vermelho). Linha sólida $f(x) = x$ indica concordância perfeita e cada ponto é a medida de um gene.

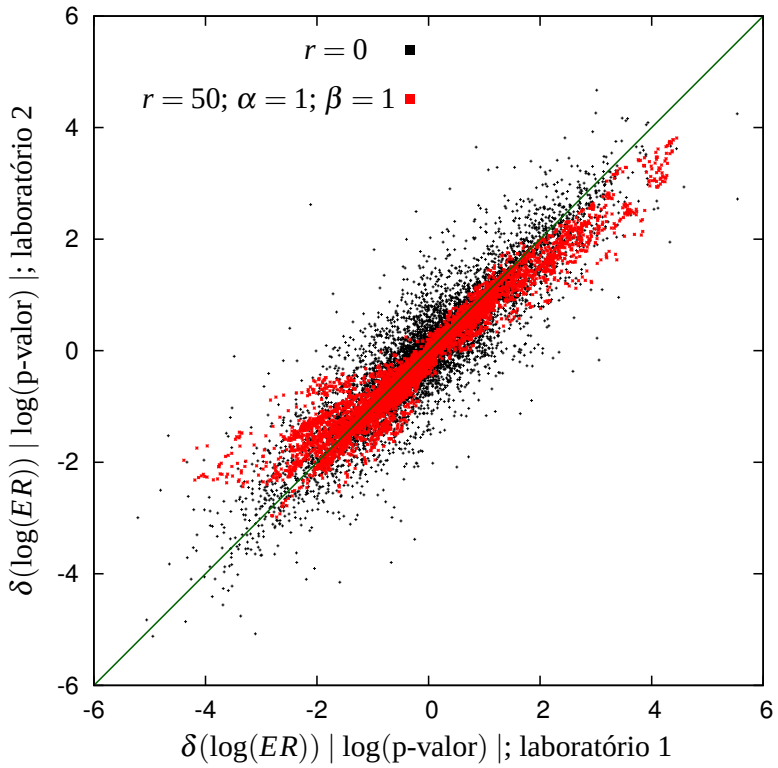


Figura 4.4: Reprodutibilidade entre laboratórios do p-valor por teste-t, para transcriptoma (preto) e transcriptograma (vermelho), sendo $\delta(x) = 1$ se $x > 0$ e $\delta(x) = -1$ se $x < 0$. Linha sólida $f(x) = x$ indica concordância perfeita e cada ponto é a medida de um gene.

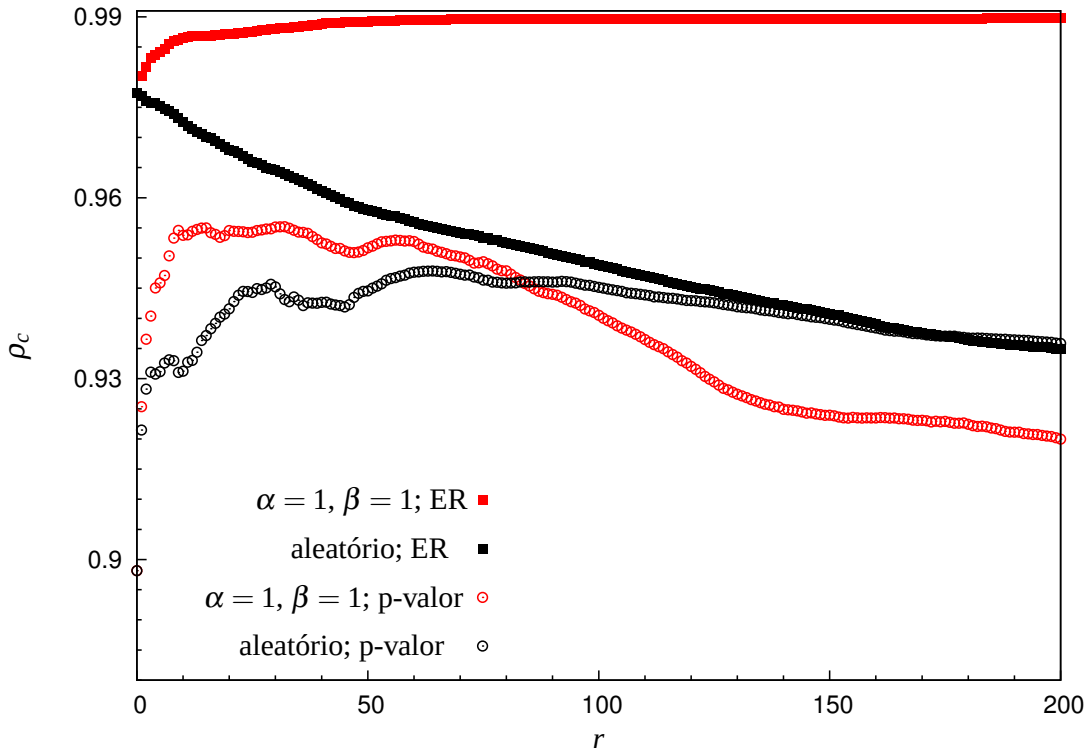


Figura 4.5: Concordância entre laboratórios

para cinco pacientes/doadores. Mostramos nas figuras 4.3, com razão de expressão entre cancêr e saudável*, e 4.4, com p-valor, como se dá a concordância dos laboratórios para todos os genes para dados de transcriptoma (transcriptograma com raio 0) e um transcriptograma, $\alpha = \beta = 1$, com raio $r = 50$.

Podemos calcular a concordância entre as medidas entre os dois laboratórios usando o coeficiente de correlação de concordância [38], ρ_c,

$$\rho_c = \frac{\sigma_{1,2}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}, \tag{4.13}$$

que mede o quanto duas medidas se aproximam da linha identidade $f(x) = x$, correspondente à concordância perfeita. Seja a população das observações, ER ou p-valor, de todos os genes, efetuada pelo laboratório i, σ_i e μ_i são respectivamente o desvio padrão e a média desta população. Por sua vez, $\sigma_{1,2}$ é a covariância entre as populações, ER ou p-valor, das observações dos laboratórios 1 e 2.

Vemos na figura 4.5 como a concordância entre os laboratórios evolui para transcriptogramas produzidos para vários raios. Transcriptograma produzidos com um ordenamento aleatório, com genes em posições aleatórios, sem minimização de energia, são mostrados também para comparação. Observa-se que o transcriptograma é efetivo

*Note, quando os dados do microarranjo são pré-processados por RMA, a expressão está em escala logarítmica e a razão da expressão entre duas condições, A e B, é dado, na verdade, por $t_i^A - t_i^B$.

ao aumentar a concordância entre laboratórios, e isto é mais um indicativo de que o transcriptograma, ao realizar médias sobre valores de expressão em genes vizinhos, está de fato reduzindo uma variabilidade indesejada da medida. Enquanto que a concordância para a expressão relativa é melhor para raios maiores do transcriptograma, observa-se que a concordância para p-valor encontra uma região de concordância otimizada aproximadamente em $10 < r < 70$. Note, a medida de expressão relativa não leva em consideração a variabilidade da medida, sendo apenas a razão da expressão média em cada grupo de amostras, sendo esta variabilidade representado na medida de p-valor. O intuito deste trabalho não é julgar qual método é o melhor para identificação de genes com expressão diferenciada, mas fica claro que o transcriptograma aumenta confiabilidade e reprodutibilidade de ambas as medidas realizadas por diferentes laboratórios.

5 *Sobre a qualidade do transcriptograma*

Vamos agora nos dedicar a investigar o problema da qualidade do transcriptograma. Este problema não pode ser abordado diretamente, pois não temos como conhecer um "bom resultado" diretamente. De fato, vamos avaliar a qualidade da informação que podemos extrair de um transcriptograma.

5.1 Informação-ruído

Aqui revisitamos a discussão sobre o ruído da medida, conforme visto na seção 4.1. Mas agora, vamos também avaliar o quanto de sinal estamos atenuando quando realizamos médias sobre valores de expressão de genes vizinhos. Seja a média global de um transcriptograma médio entre réplicas, $\langle\langle\tau_i\rangle_{rep}\rangle_N$,

$$\langle\langle\tau_i\rangle_{rep}\rangle_N = \langle\widetilde{s}\rangle_N + \langle\langle\widetilde{\xi}\rangle_{rep}\rangle_N, \tag{5.1}$$

$$\langle\langle\tau_i\rangle_{rep}\rangle_N = \langle\widetilde{s}\rangle_N, \tag{5.2}$$

onde usamos que

$$\langle\langle\widetilde{\xi}_i\rangle_{rep}\rangle_N = \langle\langle\widetilde{\xi}_i\rangle_N\rangle_{rep} \to 0, \tag{5.3}$$

por hipótese, sendo

$$\langle x\rangle_N = \frac{1}{N}\sum_{j=1}^{N} x_j. \tag{5.4}$$

com N o número total de genes da rede. Podemos medir a razão entre distância média de $\langle\tau_i\rangle_{rep}$ para sua média global e seu desvio padrão:

$$\left\langle \frac{|\langle\tau_i\rangle_{rep} - \langle\langle\tau_i\rangle_{rep}\rangle_N|}{\sigma_R(\tau_i)} \right\rangle_N = \left\langle \frac{|\widetilde{s_i} + \langle\widetilde{\xi}_i\rangle_{rep} - \langle\widetilde{s_i}\rangle_N|}{\sigma_R(\widetilde{\xi}_i)} \right\rangle_N. \tag{5.5}$$

Basicamente, estamos medindo o quanto estamos atenuando sinal em relação ao quanto estamos atenuando ruído. Por hipótese, a atenuação do sinal é independente de

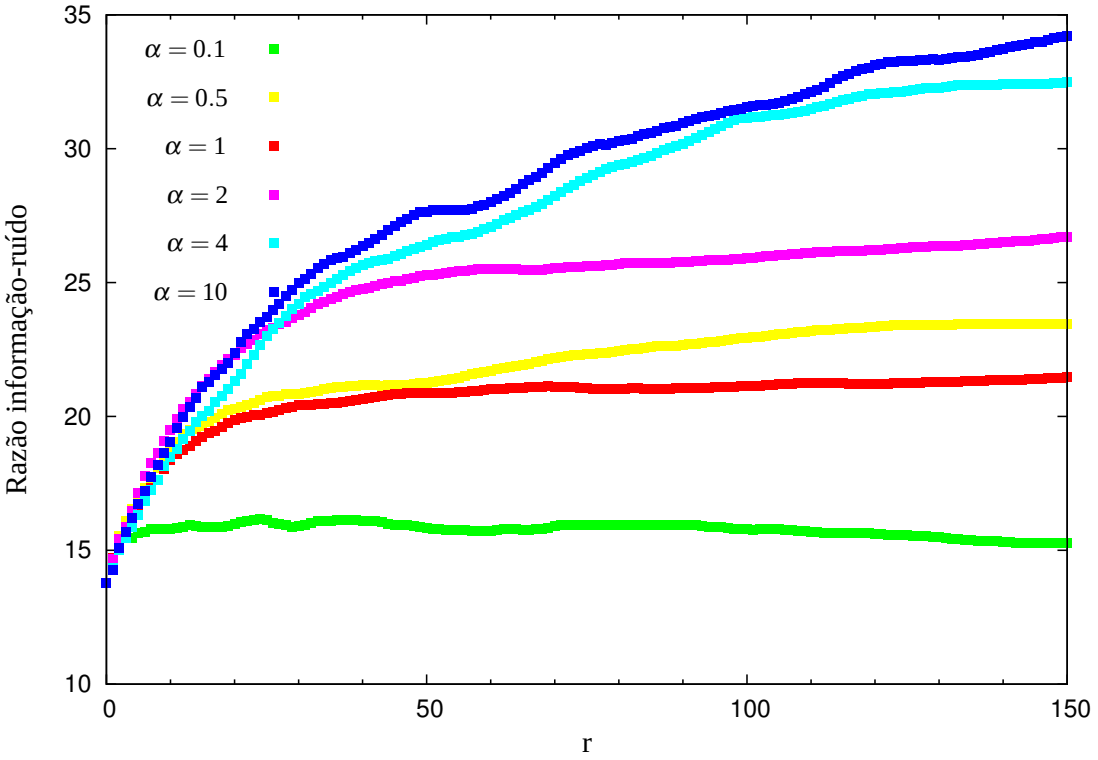


Figura 5.1: Evolução da razão informação-ruído em função do raio da janela do transcriptograma.

ordenamento, visto se tratar de ruído branco, mas a atenuação do sinal é dependente do quão forte é a correlação da expressão entre genes vizinhos. Como vemos na figura 5.1, o transcriptograma é efetivoao aumentar a razão informação-ruído, como chamaremos a medida da equação acima. Evidentemente, para raios muito grandes começamos a ter perda de informação (figura 5.2). O mais interessante destas figuras, no entanto, é que ordenamentos com valores maiores para α correspondem a melhores resultados. Isto nos parece dizer que o formato de folha da matriz de interação, onde genes ligados se encontram a uma distância máxima limite, corresponde a um transcriptograma melhor que aquele gerado com o ordenamento que até então vinha sendo usado, com $\alpha = \beta = 1$. Este favorece a formação de módulos interagentes ao custo de manter alguns pares de genes interligados a distâncias maiores. Vamos, no que segue deste capítulo, tentar corroborar este fato.

5.2 Proporcionalidade

Também parte do projeto MAQC, como os foram os dados da seção 4.2, no experimento aqui descrito foi realizado uma série de transcriptomas por microarranjo a partir de duas amostras de RNA disponíveis comercialmente. Uma destas amostras de RNA é a Referência de RNA Humano Universal (Universal Human Reference RNA,

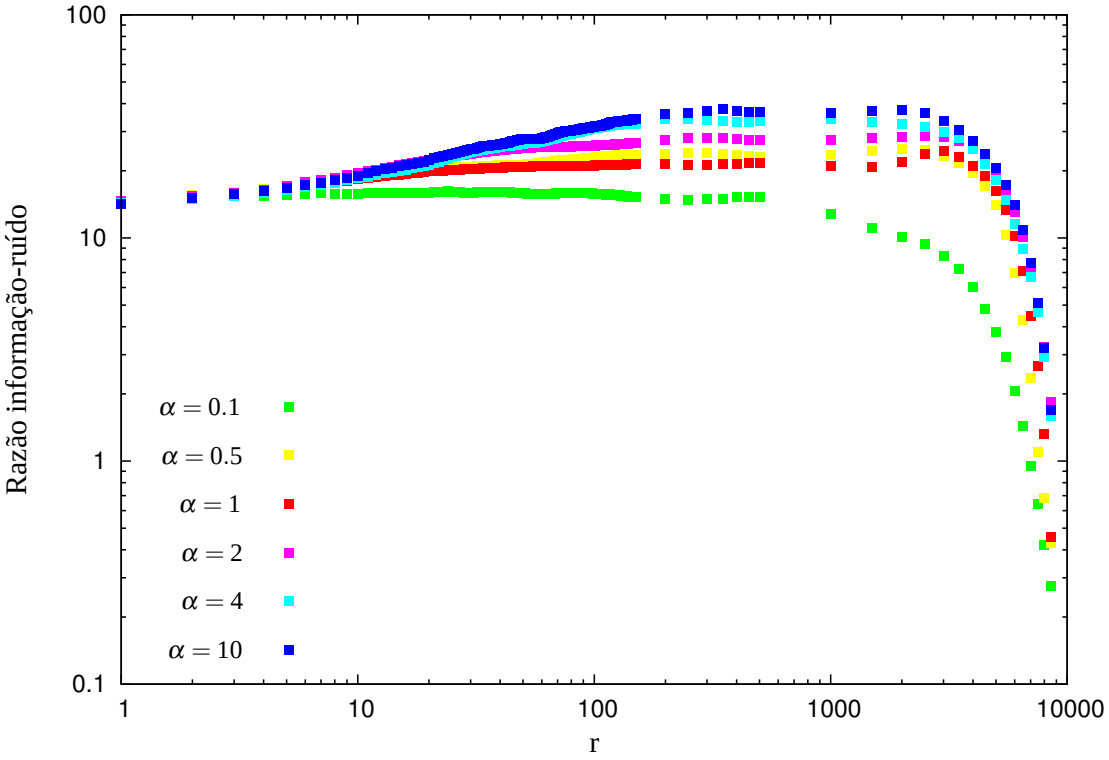


Figura 5.2: Igual a figura 5.1, mas para raio de janela até tamanho da rede.

UHRR) da empresa Stratagene, desenvolvida para servir como amostra de referência para experimentos de microarranjo de *Homo sapiens*. Esta amostra é uma mistura de RNA's de 10 linhagens de células cancerosas humanas, a saber, conforme listadas na tablea 5.1. A outra amostra de RNA usada na série de transcriptomas é a Referência de RNA de Cérebro Humano (Human Brain Reference RNA, HBRR), da empresa Ambion. Denominou-se o transcriptoma da amostra UHRR como amostra A, enquanto que o transcriptoma da amostra HBRR de amostra B. Adicionalmente, e é o que torna este experimento revelante para a discussão deste trabalho, realizaram-se transcriptomas de amostras que eram misturas destes dois RNA's: a amostra C, composta de 75% de UHRR e 25% de HBRR, e a amostra D, que por sua vez é composta de 25% de UHRR de 75% HBRR. Deste modo, a medida de expressão dos genes que são diferencialmente expressos em ambas as amostras A e B devem ter seu valor de expressão nas amostras C e D proporcionais àquelas, e a investigação deste comportamento é o objetivo desta seção.

Adenocarcinoma, glândula mamária	Melanoma
Hepatoblastoma, fígado	Lipossarcoma
Adenocarcinoma, colo do útero	Linfoma histiocitário; macrophage; histocyte
Carcinoma embrional, testículo	Leucemia linfoide, linfoblasto T
Glioblastoma, cérebro	Plasmacitoma; mieloma; linfócito B

Tabela 5.1: Linhagens cancerosas usadas para produzir a amostra UHRR

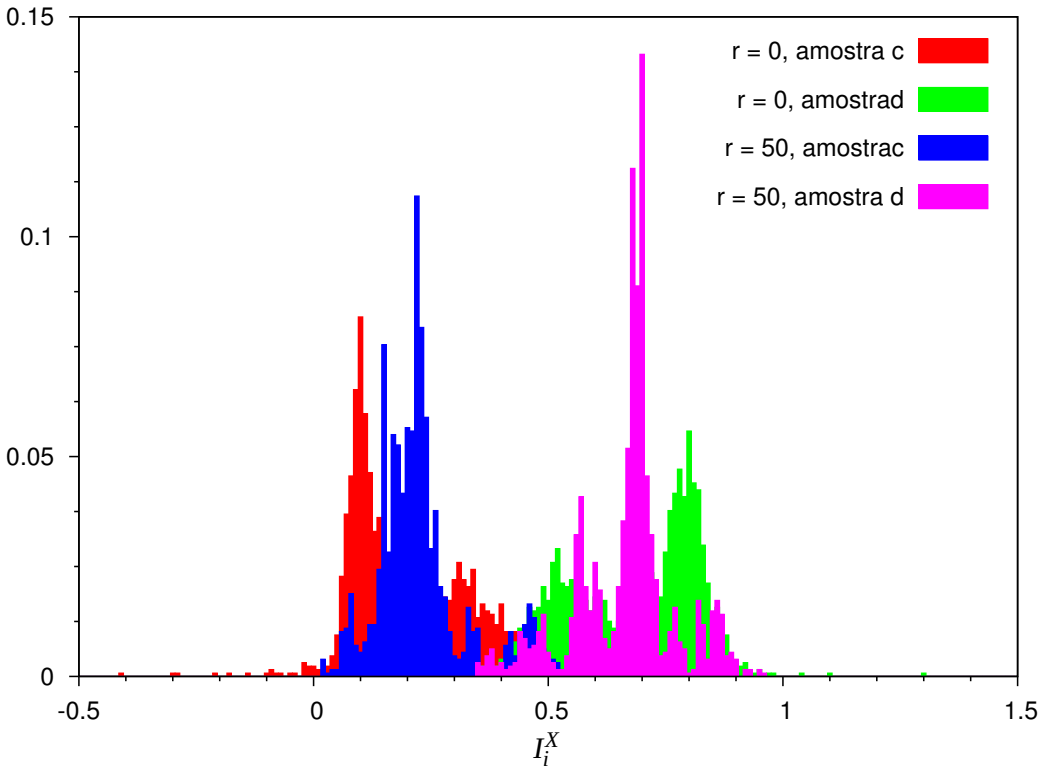


Figura 5.3: Histograma da distribuição de I_i^X, com x denotando amostra C ou D.

De modo a observar o quanto um transcriptograma respeita a proporcionalidade da expressão destas amostras com concentrações proporcionais, vamos definir um índice de intensidade relativa do sinal,

$$I_i^X = \frac{e_i^X - e_i^A}{e_i^B - e_i^A}, \tag{5.6}$$

onde e_i^Y denota a expressão do gene i na amostra Y. Observe que os dados foram pré-processados pelo método RMA, que fornece a expressão em escala logarítmica, motivo pelo qual o valor de expressão na equação acima são obtidos a partir de

$$e_i^Y = \exp(\tau_i^Y), \tag{5.7}$$

para recuperar a escala natural, sendo τ_i^Y o valor do transcriptograma no gene i para a amostra Y. Ainda, cabe dizer que só para genes diferencialmente expressos podemos calcular I_i^X, pois para genes não expressos estaríamos calculando apenas ruído. Deste modo, I_i^X só foi calculado para uma parcela dos genes, aqueles que são os 15% mais diferencialmente expressos nas amostras A e B*.

Um histograma da distribuição de I_i^X é mostrado na figura 5.3, tanto para um transcriptograma de $r = 0$, que na verdade corresponde a um transcriptoma puro (nenhum outro gene está dentro da janela de suavização do i-ésimo, além do próprio i-ésimo), e

*Diferencialmente expressos aqui estabelecido como a diferença de expressão do gene da amostra A para a amostra B.

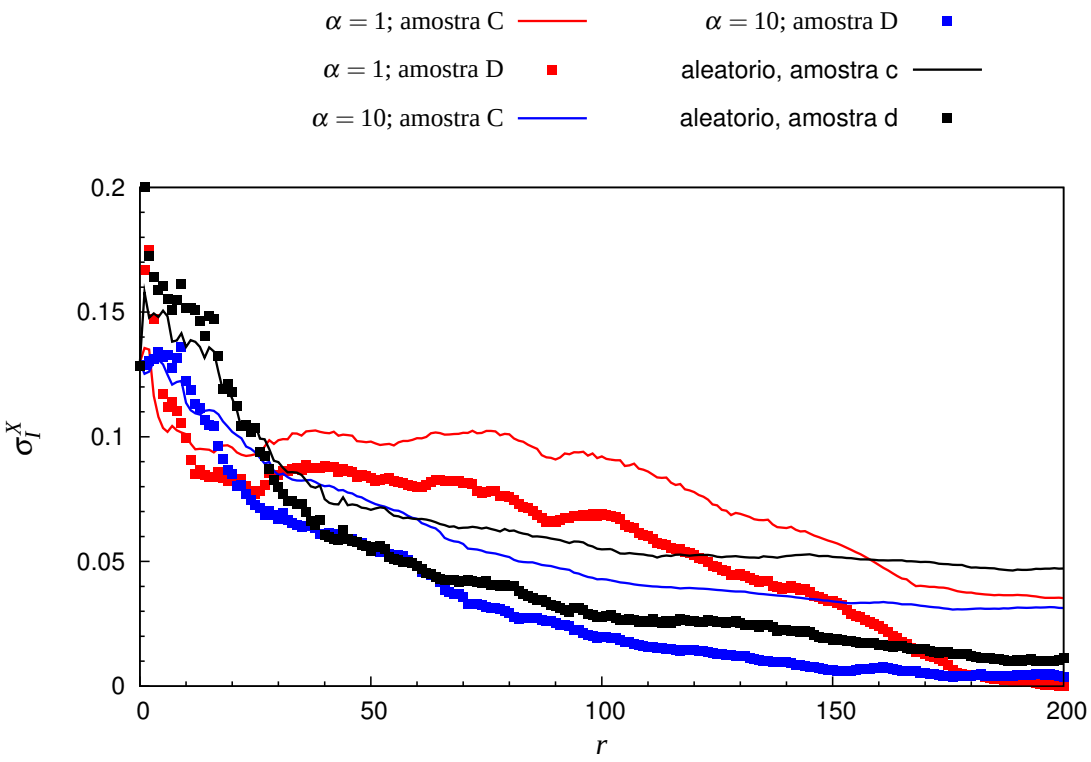


Figura 5.4: Desvio padrão da distribuição de I_i^X em função do raio da janela do transcriptograma, para as amostras $X = C$ (linha) e $X = D$ (pontos)

para um transcriptograma de raio 50 com ordenamento de $\alpha = 1$. Notamos aqui que o transcriptograma é capaz de concentrar mais a expressão dos genes das amostras C e D em torno de um ponto comum. Isto de fato é mostrado na figura 5.4, onde é mostrado o desvio padrão de I_i^X, σ_I^X, ao redor de sua média em função do raio do transcriptograma, e aqui também é mostrado para transcriptogramas produzidos para $\alpha = 10$ e com um ordenamento aleatório (genes dispostos de maneira aleatória, sem minimização de energia). Um alto valor de α é mais efetivo ao concentrar I_i^X em torno de sua média, $\overline{I^X}$. Complementarmente, mostramos a evolução de $\overline{I^X}$ na figura 5.5, onde vemos que um ordenamento aleatório, que também é capaz de diminuir o desvio da intensidade relativa, aproxima rapidamente $\overline{I^X}$ de 0.

5.3 Diagnóstico

Seguindo nossa investigação sobre a qualidade de um transcriptograma, vamos agora analisar a qualidade da informação que deste se extrai no que constitui o argumento mais relevante deste trabalho. Partindo-se de uma série de transcriptomas pertencentes a diferentes classes, pergunta-se qual a eficiência de um transcriptograma quanto a sua capacidade de distinguir entre estas. Entenda-se por classe uma característica comum a uma série de amostras, que no presente caso será a de o indivíduo de onde

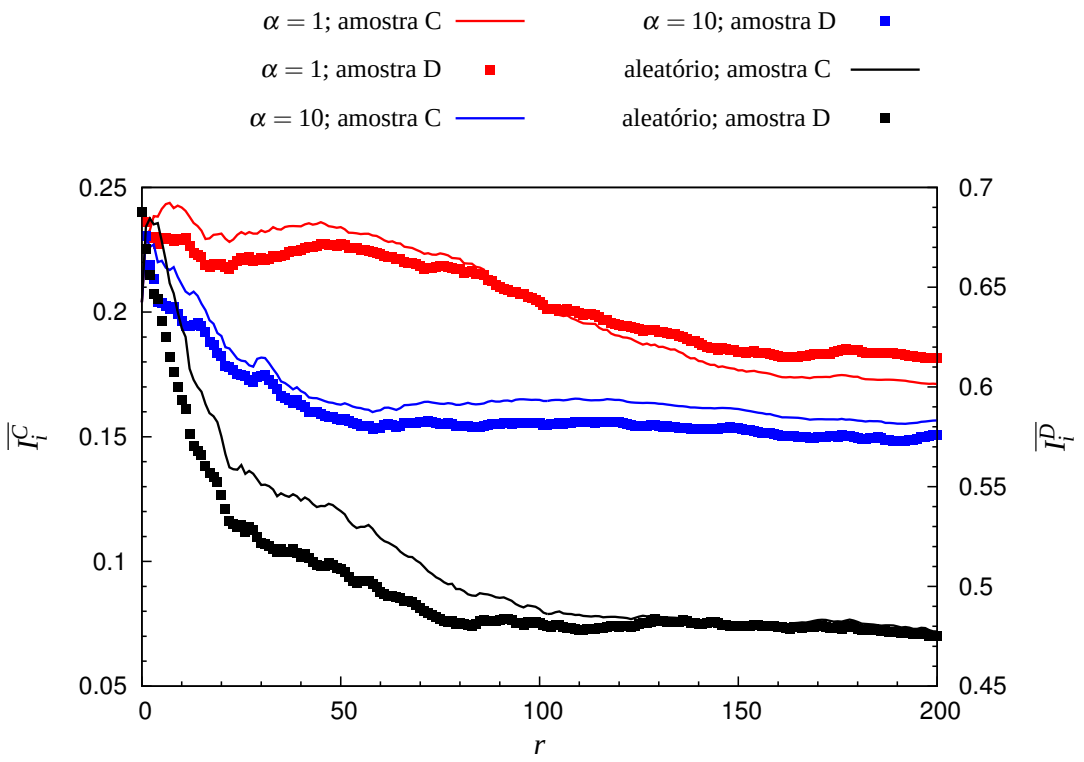


Figura 5.5: Média da distribuição de I_i^X em função do raio da janela do transcriptograma, para as amostras $X = C$ (linha) e $X = D$ (pontos)

provém a mostra apresentar ou não determinada doença. Em outras palavras, trata-se de um problema de diagnóstico, onde o objetivo é identificar uma amostra como proveniente de um indivíduo saudável ou doente.

Vamos inicialmente definir um grupo treinamento, um grupo de amostras de classe previamente conhecida a ser usada para treinar o modelo, e o grupo teste, cujas amostras serão posteriormente classificadas como pertencente a uma das classes. A partir do grupo de amostras treinamento, devemos inicialmente fazer a escolha dos classificadores, ou seja, os genes cujo valor de expressão serão usados como informação para a construção do modelo de diagnóstico. Aqui, do conjunto total de amostras, estas são divididas em 5 grupos, sendo então o diagnóstico realizado 5 vezes onde cada grupo é usado uma única vez como grupo teste enquanto os outros 4 grupos são usados como grupo treinamento. Assim, toda amostra é diagnosticada uma e apenas uma vez. A divisão dos grupos é feita de modo a respeitar a proporcionalidade entre o número de amostras de cada classe no conjunto total. A escolha de quais amostras compõem cada grupo é feito de modo aleatório, e o processo é repetido 10 vezes, visto que a eficiência do diagnóstico é dependente de quais amostras foram usadas para treinamento.

São vários os possíveis métodos de aprendizado de máquina que podemos usar para o diagnóstico. De fato, no que concerne a dados de microarranjo, não existe con-

senso quanto ao melhor método de diagnóstico. Inclusive, em anos recentes, dois desafios foram realizados à comunidade científica onde o objetivo era o de desenvolver um método de predição de classe com base em dados de microarranjo: o MAQC-II [39] [40] e o *sbvIMPROVER* [41][42]. Aos participantes dos desafios foram fornecidos um conjunto de amostras, perguntando-se qual o status destas relativo a certas doenças. Neste trabalho, não iremos nos deter sobre qual é o melhor método possível de diagnóstico. De fato, iremos apenas mimetizar o método utilizado pelo time que se sagrou vencedor do desafio sbvIMPROVER, como será descrito a seguir.

O objetivo aqui é investigar como o raio da janela do transcriptograma afeta a qualidade deste, tanto para ordenamentos produzidos com $\alpha = 1$, o ordenamento usado em estudos anteriores, quanto para $\alpha = 10$, que conforme visto na seção 5.1 aparenta produzir um sinal menos ruidoso. Entenda-se aqui por qualidade de um transcriptograma a relevância da informação biológica por ele traduzida.

A primeira etapa para a realização do diagnóstico é a escolha dos genes que melhor distinguem as classes de amostras. No modelo que aplicaremos, isto é feito escolhendo, no grupo treinamento, os dois genes com menor p-valor em um test-t de Student. Dado a natureza do transcriptograma, para raios de janela grandes é evidente que os dois genes a serem escolhidos dessa maneira serão vizinhos diretos no ordenamento, pois seus perfis de expressão são quase idênticos. Para evitar tal fato, escolhemos o gene de menor p-valor que satisfaça também a condição de estar a uma distância maior do que r, o raio da janela, do primeiro gene escolhido. Em seguida aplicamos o método de aprendizado de máquina que irá calcular um valor de confiança $p_{i,k}$ para a i-ésima amostra do grupo treinamento de pertencer a classe k, como segue.

5.3.1 LDA

Iremos agora descrever o método de aprendizado de máquina que vamos utilizar para o diagnóstico de amostras conhecido de LDA, análise de discriminante linear [43].

Pelo teorema de Bayes, podemos calcular a probabilidade posterior, $p(k \mid x)$, de uma amostra pertencer à classe k, dado que esta ocupe o ponto x no espaço de genes, com a probabilidade *a priori* da classe k, $p(k)$

$$p(k \mid x) = \frac{p(x \mid k)p(k)}{p(x)}, \tag{5.8}$$

onde o termo $p(x \mid k)$ descreve a probabilidade de uma amostra estar no ponto x do espaço de genes dado que esta pertença à classe k, chamada de densidade de classe, e $p(k)$ é a probabilidade de uma classe qualquer.

Os métodos de classificação LDA (linear discriminant analysis) e QDA (quadratic discriminant analysis) surgem quando se assume que a densidade da classe k é uma distribuição gaussiana,

$$p(x \mid k) = c_k \, exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right], \tag{5.9}$$

onde Σ_k é a matriz de covariância, μ_k é o centro da distribuição e c_k é a constante de normalização,

$$c_k = \frac{1}{(2\pi)^{g/2} \mid \Sigma_k \mid^{1/2}}, \tag{5.10}$$

sendo g o número de dimensões, ou genes, do sistema. A partir da equação 5.8, pode-se procurar pelos hiperplanos onde duas classes são igualmente prováveis. Tais hiperplanos dividem os espaço de genes nas regiões pertencentes à cada classe, classificando assim como pertencente à classe k as amostras localizadas na região onde a probabilidade posterior desta classe é igual à probabilidade de todas as outras classes. Por exemplo, para o caso de duas classes, a e b, podemos escrever o hiperplano de igual probabilidade posterior para ambas a partir do logaritmo da razão destas:

$$\begin{aligned}
0 =& ln\left(\frac{p(a \mid x)}{p(b \mid x)}\right) = ln\left(\frac{p(x \mid a)p(a)}{p(x \mid b)p(b)}\right) \\
=& ln\left(\frac{c_a p(a)}{c_b p(b)}\right) - \frac{1}{2}\left[(x-\mu_a)^T \Sigma_a^{-1}(x-\mu_a) - (x-\mu_b)^T \Sigma_b^{-1}(x-\mu_b)\right] \\
=& x^T\left(\Sigma_a^{-1} - \Sigma_b^{-1}\right)x + 2\left(\mu_a^T \Sigma_a^{-1} - \mu_b^T \Sigma_b^{-1}\right)x + \\
&+ \mu_a^T \Sigma_a^{-1}\mu_a - \mu_b^T \Sigma_b^{-1}\mu_b - 2\, ln\left(\frac{c_a p(a)}{c_b p(b)}\right).
\end{aligned} \tag{5.11}$$

A equação acima é quadrática em x para descrever o hiperplano de igual probabilidade, o que da nome ao método QDA. O método LDA, por sua vez, surge ao se impor a hipótese ainda mais forte de que as distribuições normais que descrevem $p(x \mid k)$ têm igual matriz de covariância para todas as classes,

$$\Sigma_a = \Sigma_b = \Sigma,$$

o que elimina o termo quadrático na equação 5.11, além de eliminar os fatores de normalização c_k:

$$2(\mu_a - \mu_b)^T \Sigma^{-1} x + (\mu_a + \mu_b)\Sigma^{-1}(\mu_a - \mu_b) - 2\, ln\left(\frac{p(a)}{p(b)}\right) = 0 \tag{5.12}$$

Esta equação, agora linear em x, descreve o hiperplano usado para classificação no método LDA. Uma nova amostra i é então classificada, pela coordenada x em que se

localiza no espaço de classificadores, como pertencente a classe a se

$$(\mu_a - \mu_b)^T \Sigma^{-1} x > -\frac{1}{2}(\mu_a + \mu_b)\Sigma^{-1}(\mu_a - \mu_b)\ ln\left(\frac{p(a)}{p(b)}\right), \tag{5.13}$$

com confianças $p_{i,a} = 1$ e $p_{i,b} = 0$, e amostra da classe b caso contrário, com confianças $p_{i,a} = 0$ e $p_{i,b} = 1$. Os parâmetros μ_k, Σ e $p(k)$ estimados a partir do grupo treinamento.

O diagnóstico foi implementado usando o pacote MLInterfaces [44], pacote integrante do projeto Bioconductor [25], no programa estatístico R [26].

5.3.2 Eficiência do diagnóstico

Vários são os possíveis métodos que podem ser usados para se avaliar a qualidade de um diagnóstico. Aqui apresentamos o CCEM (*correct class enrichment metric*), que para N_t amostras do grupo teste é dado por

$$\text{CCEM'} = \sum_i p_{i,c(i)}\delta_i, \tag{5.14}$$

onde $c(i)$ indica a classe a qual a amostra i verdadeiramente pertence, $p_{i,k}$ indica a confiança da predição da amostra i pertencer a classe k e

$$\delta_i = \begin{cases} 1 & \text{se amostra} i corretamente classificada \\ -1 & \text{se amostra} i incorretamente classificada \end{cases}. \tag{5.15}$$

O CCEM' medirá a eficiência do diagnóstico com valores entre $-N$, caso todas as amostras sejam classificadas incorretamente com confiança 1, e N, caso todas sejam corretamente classificadas com confiança 1. Normalizando,

$$CCEM = \frac{(\frac{CCEM'}{N} + 1)}{2}, \tag{5.16}$$

obtém-se uma eficiência que varia de 0 a 1.

5.3.3 Psoríase

Psoríase é uma doença autoimune, inflamatória crônica de pele. Em muitos casos, é acompanhada de artrite inflamatória e seu diagnóstico tipicamente se dá por análise física da pele lesionada. Foram utilizados amostras de 2 experimentos, ambos cadastrados disponíveis no GEO sobre o código GSE13355 [5] [6] e GSE14905 [45]. Juntos, estes experimentos possuem 91 amostras de pacientes com psoríase, tecido lesionado, e 85 amostras de doadores saudáveis, todas amostras de pele.

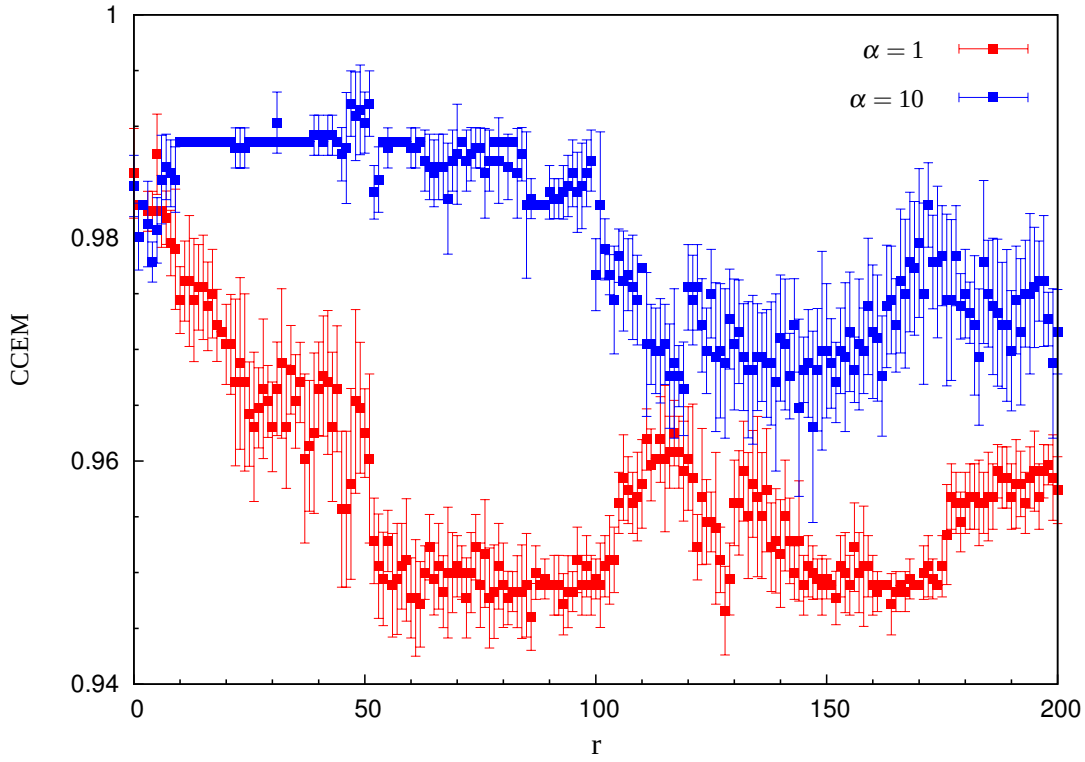


Figura 5.6: Eficiência do diagnóstico de psoríase em função do raio r da janela, ambos os ordenamentos com $\beta = 1$.

A figura 5.6 mostra a eficiência, segundo CCEM, para o diagnóstico de psoríase para ordenamentos produzidos com os parâmetros $\alpha = 1$, $\beta = 1$ e $\alpha = 10$, $\beta = 1$. Nota-se capacidade superior do ordenamento em forma de folha, $\alpha = 10$, em gerar valores de expressão capazes de distinguir entre os estados doente/saudável. Note, para $r = 0$ estamos falando de dados não janelados, ou seja, dados diretamente retirados de transcriptoma sem médias sobre vizinhança.

5.3.4 Esclerose múltipla

A esclerose múltipla é uma doença autoimune que afeta o sistema nervoso central. As amostras foram obtidas de [46], experimento cadastrado no ArrayExpress sob o código E-MATB-69, com 26 amostras de pacientes com esclerose múltipla, sendo 12 em fase recidiva e 14 em fase remitente, e 18 amostras controle. Aqui, todas as amostras são de linfócitos de sangue periférico, e as amostras controle são retiradas de pacientes com outras desordens neurológicas de caráter não-inflamatório.

Mais uma vez, vemos nas figuras 5.7, onde realizamos um diagnóstico para diferenciar doadores saudáveis de pacientes doentes, como na figura 5.8, onde realizamos um diagnóstico para diferenciar os pacientes em faze recidiva e remitente, que o ordenamento para α alto é muito superior para a realização do diagnóstico. Para esta última

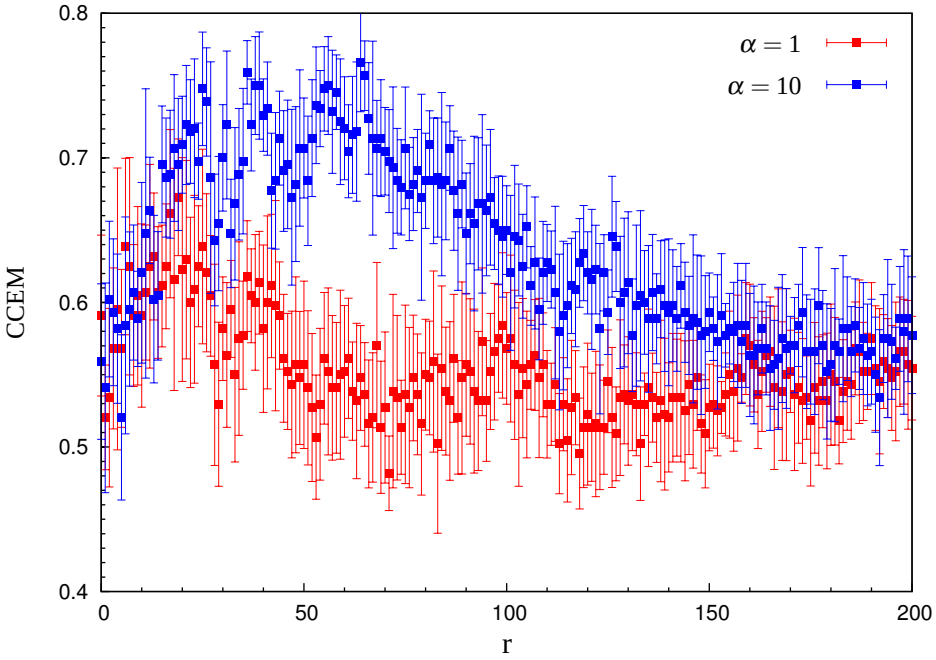


Figura 5.7: Eficiência do diagnóstico de esclerose múltipla em função do raio r da janela, ambos os ordenamentos com $\beta = 1$.

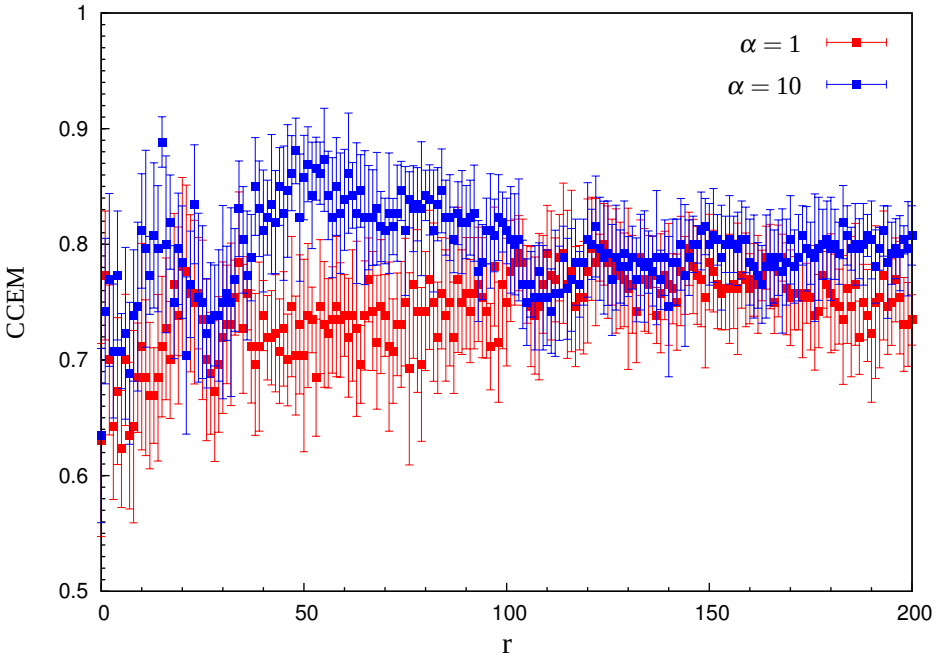


Figura 5.8: Eficiência do diagnóstico das fases da esclerose múltipla, recidiva e remitente, em função do raio r da janela, ambos os ordenamentos com $\beta = 1$.

Tabela 5.2: Eficiência do diagnóstico por CCEM para o time vencedor do desfio *sbvIMPROVER*, transcriptograma de raio zero e transcriptograma de $\alpha = 10$ no raio onde eficiência foi máxima

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

figura, ainda, ressalta-se que o aumento da eficiência em transcriptograma de raio $r = 0$ para $r \sim 50$ foi bastante significativa.

Observamos aqui que para as duas doenças que analisamos o transcriptograma é efetivo ao aumentar a eficiência do diagnóstico. Podemos comparar nosso desempenho com o do time vencedor do desafio *sbvIMPROVER*, como vemos na tabela 5.2. Observe, o diagnóstico usando transcriptograma pressupõe uma seleção prévia dos genes, visto que o ordenamento é feito com aquelas proteínas que apresentam pelo menos uma interação listada no STRING com *score* de confiança maior que 0.8. O desafio de distinguir entre estágios da esclerose múltipla foi retirado do desafio por não haver número suficiente de times com eficiência superior a um limite mínimo imposto pelos organizadores. Nota-se, pela tabela 5.2 e figuras 5.6, 5.8 e 5.7, que um transcriptograma com ordenamento em formato de folha, produzido com $\alpha = 10$, é seguramente capaz de produzir informação com relevante significado biológico, melhorando os dados de transcriptoma medidos por microarranjo.

6 Conclusões

Este trabalho se dedicou a pesquisar em detalhes um método, o transcriptograma, para tratamento de dados de uma medida biológica, o microarranjo, que visa ler a expressão gênica de um organismo. A dificuldade primária que se impõe é a de que não podemos conhecer a medida ideal, o estado verdadeiro e sem ruído. Apesar de haver trabalhos que comparam medidas de microarranjo com outra técnica quantitativa de expressão gênica, por exemplo [20], estes se limitam a analisar um conjunto muito restrito de genes, no máximo uma centena, enquanto o transcriptograma deve necessariamente levar em consideração o genoma completo. Inclusive, considerando que fazemos médias sobre a expressão em um grupo de genes, não podemos compará-lo com a expressão de um único gene. Segue, portanto, que devemos avaliar o transcriptograma não com base no quanto este aproxima o perfil de expressão obtido com o verdadeiro, o que é impossível, mas sim sobre a qualidade da informação de que dele se extrai.

É sobre esta ótica que o resultado mais relevante deste trabalho, sem dúvida, é o de que o transcriptograma é capaz de melhorar substancialmente a eficiência de um modelo preditivo, aqui demonstrado para os diagnósticos de psoríase e esclerose múltipla. Inclusive, a eficiência máxima obtida é superior ao do vencedor do desafio sbvIMPROVER nas duas doenças aqui estudadas. Tão ou mais notável ainda é o de que esta qualidade superior da informação retirada do transcriptograma se dá para um conjunto de parâmetros, $\alpha = 10$ e $\beta = 1$, que não era o até então utilizado em qualquer trabalho contendo transcriptograma ([1], [28], [47], [29]), sendo $\alpha = 1$ e $\beta = 1$ o único ordenamento considerado. Conforme mostrado, o ordenamento obtido com $\alpha = 10$, em formato de folha, é menos eficiente em agrupar genes com funções biológicas em comum. O porquê de este ordenamento ser superior é, diga-se, uma pergunta em aberto. É evidente que uma investigação para mais valores de α e β é necessário, mas é difícil dizer se há um valor para α e β otimizado em geral, como para diagnóstico de todas as doenças possíveis. Mais provável seria de que o "melhor" dependa do problema em questão.

O provável próximo passo na aplicação do transcriptograma será o de um aprofundamento na questão da reprodutibilidade da análise. A consistência dos genes identi-

ficados como diferencialmente expressos em diferentes estados constitui preocupação atual [48], referência na qual, com amostras de psoríase, está demonstrado que a concordância entre experimentos realizados com a mesma plataforma é maior do que com plataformas diferentes. Tal fato deve, necessariamente, ser investigado sob a ótica do transcriptograma. O que demonstramos sobre a questão neste trabalho já é por si só bastante relevante, sendo transcriptograma capaz de aumentar a reprodutibilidade de resultabos obtidos em medidas de microarranjo quando uma mesma amostra é analisada por laboratórios diferentes.

Existem também outras possibilidades em abertas, como o uso de um outro banco de dados, o STITCH (Search Tool for Interactions of Chemicals, [49]), projeto irmão ao STRING, que lista as interações entre proteínas e substâncias químicas. É uma possibilidade investigar as modificações no transcriptograma ao se levar tais interações em consideração para ordenar a lista de proteínas. Pode-se citar também a possibilidade de se montar a matriz usando interações não booleanas, ou seja, que diferentes interações tenham pesos diferentes, como baseadas no score de confiança do STRING para a interação seja verdadeira. Poderia-se usar, ainda, ligações não direcionadas.

Adicionalmente a investigação da eficiência do transcriptograma da qual trata este trabalho, estamos desenvolvendo um programa, a ser disponibilizado para a comunidade científica, com interface gráfica para a produção de ordenamentos e trasncriptogramas. A lista de funções do programa que estamos desenvolvendo são:

- Produzir um ordenamento a partir de uma rede de interação entre proteínas inserida pelo usuário;
- Calcular as propriedades do ordenamento: perfis de modularidade, conectividade e clusterização;
- Calcular as propriedades da rede: distribuição de conectividade $p(k)$, assortatividade $S(k)$ e a clusterização média de todos os genes com conectividade k, $C(k)$;
- Calcular o número de pares de proteínas interagentes em função da distância d em que estes se encontram no ordenamento;
- Calcular transcriptogramas;
- Calcular perfis de enriquecimento funcional de funções biológicas;
- Calcular médias e desvios de transcriptogramas;
- Produzir um clustering hierárquico das amostras de transcriptograma.

O programa encontra-se em fase final de desenvolvimento, e uma versão preliminar do manual deste encontra-se no apêndice C.

Apêndices

A Estimativa da média por Tukey's Biweigth

No método Tukey Biweigth, o objetivo é calcular a média de um conjunto de dados impedindo que pontos que fujam da distribuição alterem muito o resultado. Inicialmente, calcula-se a mediana da distribuição, M, e então a mediana das distâncias absolutas até M, S. Para cada valor x_i da distribuição, calcula-se

$$u_i = \frac{x_i - M}{cS + \varepsilon}, \tag{A.1}$$

onde c é uma constante de ajuste e ε impede divisão por zero. Por padrão do método MAS5 descrito neste trabalho, usa-se $c = 5$ and $\varepsilon = 0.0001$.

Finalmente, a estimativa da média da distribuição é dada por

$$w_i = \begin{cases} (1-u^2)^2, & \mid u \mid \leqslant 0 \\ 0, & \mid u \mid > 0 \end{cases} \tag{A.2}$$

$$\mathrm{TB}_i(x_i) = \frac{\sum_i w_i x_i}{\sum_i w_i} \tag{A.3}$$

B Ontologias

Tabela B.1: Ontologias

Código	Nome
GO:0000082	G1/S transition of mitotic cell cycle
GO:0000122	negative regulation of transcription from RNA polymerase II promoter
GO:0000226	microtubule cytoskeleton organization
GO:0000398	mRNA splicing, via spliceosome
GO:0001503	ossification
GO:0001756	somitogenesis
GO:0001817	regulation of cytokine production
GO:0001932	regulation of protein phosphorylation
GO:0002237	response to molecule of bacterial origin
GO:0002700	regulation of production of molecular mediator of immune response
GO:0002703	regulation of leukocyte mediated immunity
GO:0002819	regulation of adaptive immune response
GO:0003002	regionalization
GO:0003012	muscle system process
GO:0006000	fructose metabolic process
GO:0006082	organic acid metabolic process
GO:0006091	generation of precursor metabolites and energy
GO:0006213	pyrimidine nucleoside metabolic process
GO:0006220	pyrimidine nucleotide metabolic process
GO:0006260	DNA replication
GO:0006270	DNA replication initiation
GO:0006284	base-excision repair
GO:0006289	nucleotide-excision repair
GO:0006351	transcription, DNA-dependent
GO:0006357	regulation of transcription from RNA polymerase II promoter

Continua na próxima página

Tabela B.1 – continuação da página anterior

Código	Nome
GO:0006364	rRNA processing
GO:0006401	RNA catabolic process
GO:0006412	translation
GO:0006413	translational initiation
GO:0006414	translational elongation
GO:0006520	cellular amino acid metabolic process
GO:0006754	ATP biosynthetic process
GO:0006813	Potassium ion transport
GO:0006816	calcium ion transport
GO:0006915	apoptotic process
GO:0006936	muscle contraction
GO:0006937	regulation of muscle contraction
GO:0007015	actin filament organization
GO:0007059	chromosome segregation
GO:0007067	mitosis
GO:0007167	enzyme linked receptor protein signaling pathway
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway
GO:0007179	transforming growth factor beta receptor signaling pathway
GO:0007219	Notch signaling pathway
GO:0007229	integrin-mediated signaling pathway
GO:0007243	intracellular protein kinase cascade
GO:0007249	I-kappaB kinase/NF-kappaB cascade
GO:0007254	JNK cascade
GO:0007265	Ras protein signal transduction
GO:0007265	Ras protein signal transduction
GO:0007266	Rho protein signal transduction
GO:0007608	sensory perception of smell
GO:0008380	RNA splicing
GO:0008543	fibroblast growth factor receptor signaling pathway
GO:0008624	apoptotic signaling pathway
GO:0009116	nucleoside metabolic process
GO:0009117	nucleotide metabolic process

Continua na próxima página

Tabela B.1 – continuação da página anterior

Código	Nome
GO:0009123	nucleoside monophosphate metabolic process
GO:0009165	nucleotide biosynthetic process
GO:0009190	cyclic nucleotide biosynthetic process
GO:0009309	amine biosynthetic process
GO:0009617	response to bacterium
GO:0009755	hormone-mediated signaling pathway
GO:0010551	regulation of transcription from RNA polymerase II promoter
GO:0010553	negative regulation of transcription from RNA polymerase II promoter
GO:0010629	negative regulation of gene expression
GO:0010740	positive regulation of intracellular protein kinase cascade
GO:0015031	protein transport
GO:0015931	nucleobase-containing compound transport
GO:0015986	ATP synthesis coupled proton transport
GO:0016054	organic acid catabolic process
GO:0016055	Wnt receptor signaling pathway
GO:0016338	calcium-independent cell-cell adhesion
GO:0016458	gene silencing
GO:0016567	protein ubiquitination
GO:0016575	histone deacetylation
GO:0018108	peptidyl-tyrosine phosphorylation
GO:0019221	cytokine-mediated signaling pathway
GO:0022618	ribonucleoprotein complex assembly
GO:0022900	electron transport chain
GO:0030029	actin filament-based process
GO:0030036	actin cytoskeleton organization
GO:0030098	lymphocyte differentiation
GO:0030155	regulation of cell adhesion
GO:0030198	extracellular matrix organization
GO:0030334	regulation of cell migration
GO:0030501	positive regulation of bone mineralization
GO:0030509	BMP signaling pathway
GO:0031098	stress-activated protein kinase signaling cascade
GO:0031396	regulation of protein ubiquitination

Continua na próxima página

Tabela B.1 – continuação da página anterior

Código	Nome
GO:0031589	positive regulation of locomotion
GO:0032582	negative regulation of transcription, DNA-dependent
GO:0032583	regulation of transcription, DNA-dependent
GO:0034097	response to cytokine stimulus
GO:0034220	Ion transmembrane transport
GO:0034655	nucleobase-containing compound catabolic process
GO:0035023	regulation of Rho protein signal transduction
GO:0040017	cell-substrate adhesion
GO:0042035	regulation of cytokine biosynthetic process
GO:0042110	T cell activation
GO:0042254	ribosome biogenesis
GO:0042327	positive regulation of phosphorylationset output P10b.epsP100042509
GO:0042509	regulation of tyrosine phosphorylation of STAT protein
GO:0042773	ATP synthesis coupled electron transport
GO:0042981	regulation of apoptotic process
GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB cascade
GO:0043388	positive regulation of DNA binding
GO:0043408	regulation of MAPK cascade
GO:0044262	cellular carbohydrate metabolic process
GO:0044270	cellular nitrogen compound catabolic process
GO:0044275	cellular carbohydrate catabolic process
GO:0045165	cell fate commitment
GO:0045333	cellular respiration
GO:0045667	regulation of osteoblast differentiation
GO:0045892	negative regulation of transcription, DNA-dependent
GO:0046131	pyrimidine ribonucleoside metabolic process
GO:0046649	lymphocyte activation
GO:0048193	Golgi vesicle transport
GO:0048660	regulation of smooth muscle cell proliferation
GO:0050670	regulation of lymphocyte proliferation
GO:0050730	regulation of peptidyl-tyrosine phosphorylation
GO:0050817	coagulation
GO:0050867	positive regulation of cell activation

Continua na próxima página

Tabela B.1 – continuação da página anterior

Código	Nome
GO:0050890	cognition
GO:0051028	mRNA transport
GO:0051056	regulation of small GTPase mediated signal transduction
GO:0051090	regulation of sequence-specific DNA binding transcription factor activity
GO:0051092	positive regulation of NF-kappaB transcription factor activity
GO:0051099	positive regulation of binding
GO:0051168	nuclear export
GO:0051169	nuclear transport
GO:0051258	protein polymerization
GO:0051302	regulation of cell division
GO:0051438	regulation of ubiquitin-protein ligase activity
GO:0051707	response to other organism
GO:0055002	striated muscle cell development
GO:0060047	heart contraction
GO:0060393	regulation of pathway-restricted SMAD protein phosphorylation
GO:0060415	muscle tissue morphogenesis
GO:0070085	Glycosylation
GO:0070663	regulation of leukocyte proliferation
GO:0070665	positive regulation of leukocyte proliferation

C Transcriptogramer: manual

The Transcriptogramer

Software for Transcriptogram production and gene network analysis

Reference manual

by

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

Instituto de Física - Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves 9500 - Caixa Postal 15051
91501-970 Porto Alegre, RS, Brazil

Welcome to The Transcriptogramer V.1.0

The Transcriptogramer is a tool intended to analyze gene networks and produce transcriptograms from genome wide gene expression datasets. It is released under an open source license and is completely free for any use.

The motivation for this software comes from three previous publications:

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., Towards a genome-wide transcriptogram: the *Saccharomyces cerevisiae* case. *Nucleic Acids Res.*, 39, 3005-3016 (2011).
PMID:21169199
2. Perrone, G.C., da Silva, S.R., de Almeida, R.M.C. Transcriptograms in two dimensions. To appear.
3. da Silva, S.R, Perrone, G.C., de Almeida R.MC. Transcriptograms strategies and statistics. To appear.

The first paper introduces transcriptograms as a method to analyze transcriptomic data, with expression levels projected on a 1D gene list, arranged such that the probability that gene products participate in the same metabolic pathway exponentially decreases with distance between the genes on the list. Averages over expression levels of neighboring genes strongly reduce the relevance of the typical noise in microarray measurements. Transcriptograms are hence genome wide gene expression profiles that provide a global view for the cellular metabolism, while indicating gene sets whose expression are altered. The second paper proposes a two dimensional gene arrangement, and shows that a second dimension may improve still further noise reduction by allowing a less frustrated gene arrangement. Finally, the third paper provides case studies in 1D and 2D, suggesting possible uses for the transcriptogram and the appropriate statistical analyses.

See also

How to use this document

This document provides information on how to get started with The Transcriptogramer, how the software operates, how to analyze the results.

It assumes you are familiar with genomic, protein-protein interaction networks, transcriptomic and proteomic data generated by high-throughput experiments. You should also be familiar with fundamentals of graph theory in its applications to gene/protein networks.

Getting started

The Transcriptogramer works on an association matrix to produce an ordered arrangement (in one or two dimensions) of genes such that the distance between the genes on the arrangements correlates with the probability that the genes are associated, as informed by the association matrix. The correlation exponentially decays with the distance. When some attribute is assigned to each genes (for example, the gene expression level), it is possible to produce a profile by plotting the attribute value of each gene versus gene position in the gene arrangement. Usually these attributes may strongly vary from one gene to its neighbors, causing the profiles to fluctuate wildly.

On the other, supposing the attributes are such that there is correlation between their values if the genes are associated (as informed by the association matrix), arranging the genes by their association probability could yield a smoother profile (neighboring genes should present similar attribute values).

Transcription levels as measured with microarrays are very noisy, meaning that the levels have two important components: a signal and a noise. When arranging the genes on a list or on a plane by the probability their products are associated in a metabolic pathway, for example, will cause the signal component of transcription level measurements to be correlated, but will not have the same effect on the noise component, which is expected to be non-correlated.

Transcriptograms uses both noise lack of correlation for any pair of genes **and** signal correlation between nearby genes on an adequate arrangement to produce smoother transcription level profiles. This is achieved by averaging transcription levels on a given neighborhood (window) of each gene: the average should keep information on correlated signals, but strongly reduce uncorrelated, white noise. These profiles may then be a tool to either globally assesses metabolic states as to identify which modules or pathways present correlated alterations.

There is a necessary workflow to produce transcriptograms using The Transcriptogramer: i) provide an association matrix, ii) order the genes/proteins by the probability that they are associated, iii) link the transcription dataset to each gene of the ordering and calculate the averages over the chosen window. The final product, the transcriptogram, is a table that gives the value of the averaged transcription level as a function of the gene position on the arrangement. With this table it is a straightforward step to plot the profiles by using your favorite plot applicative.

Transcriptograms need biological interpretation, as the plots will indicate the metabolic altered regions, but will not name it in the general case. For *Homo sapiens*, The Transcriptogramer package provides biological characterization of the regions of the gene ordering, based on terms of the Gene ontology: Biological Process (GO:BP) [GeneOntology2000]. However, for other, customized gene sets or organisms you have to do it. A possible way to biological characterize the gene arrangements is by projecting GO:BP terms, metabolic pathway gene list or a customized gene set on the ordered gene arrangement. This is done by first choosing a gene set and then assigning to each position on the ordering the values 1 or 0 depending on whether the gene on that position respectively belongs or not to the gene set. Performing an average over a window (neighborhood) of each gene, produces a profile giving the probability density that in a given region of the ordering there are genes that belong to the chosen gene set. The Transcriptogramer calculates these density profiles.

For each step of the workflow, there is an appropriate tab in The Transcriptogramer, as we explain in what follows, in the Tabs and Functions section.

TAB 1 - Orderer

This step produces arrangement in one or two dimensions such that the probability that two genes/proteins are associated exponentially decays with the distance between their positions on the arrangement.

Figure 1a presents The Transcriptogramer at TAB1, intended to arrange the genes/proteins such that the probability that they are associated exponentially decays with the distance between their positions on the arrangement. This is achieved, as prescribed in Ref. [Rybarczyck2011] by minimizing a cost function F, defined as

$$F = \sum_{\vec{p}} A(\vec{p},\vec{q})|\vec{p}-\vec{q}|^{\alpha} \sum_{\vec{v}} (|A(\vec{p},\vec{q}) - A(\vec{p},\vec{q}+\vec{v})| + |A(\vec{p},\vec{q}) - A(\vec{p},\vec{q}-\vec{v})| + |A(\vec{p},\vec{q}) - A(\vec{p}+\vec{v},\vec{q})| + |A(\vec{p},\vec{q}) - A(\vec{p}-\vec{v},\vec{q})|), \quad (1)$$

where $\vec{p}$ and $\vec{q}$ are possible positions for the genes in the orderings. In 1 dimension, for a list of N genes, $\vec{p}$ and $\vec{q}$ are the integers in the interval $[1, N]$. In two dimensions, $\vec{p}$ and $\vec{q}$ are the possible positions on a plane, being two dimensional vectors. $A(\vec{p},\vec{q})$ is a matrix whose elements assume the value 1 when the genes/proteins located at $\vec{p}$ and $\vec{q}$ are associated and assume the null value otherwise. The sum over $\vec{v}$ stands for a sum over the neighbors of $\vec{p}$ and $\vec{q}$. α is a free parameter, whose role is to balance the contribution from the first factor $|\vec{p}-\vec{q}|^{\alpha}$, which increases with the distance on the ordering between two associated genes/proteins, and the second factor, represented by the sum over the neighbors of $\vec{p}$ and $\vec{q}$, which favors that these associate genes/proteins, are associated to other genes/proteins that are associated between themselves.

The ordering algorithm starts by randomly arranging the genes/proteins listed in the input association file described in the previous section. The cost function F_i for this initial configuration is calculated as described in Eq. (1). Two positions are then randomly chosen and their content is swapped (in two dimensions a location may be empty) and a new value for the cost function F_f is calculated. If $\Delta F = F_f - F_i < 0$, the swap is accepted and a new pair of positions is randomly chosen to repeat the process. If $\Delta F = F_f - F_i = 0$, the swap is accepted with 50% of probability, and finally if $\Delta F = F_f - F_i > 0$, the swap is accepted with probability of $e^{-\frac{\Delta F}{T}}$, where T is a parameter that allows some swappings that increase the cost function, analogous to the temperature that allows a system to explore some states with higher energy in a Monte Carlo simulation of a physical system [Metropolis1949]. To avoid metastable states, we minimized the cost function following a method known as simulated annealing [Kirkpatrick1983]: the temperature is initially set at high values and at arbitrary intervals, is reduced up to very low values.

The inputs for this step are i) the association matrix, which could be the output of the step performed in TAB1, or produced elsewhere, provided it is a plain file with two columns, with each row informing the associated pairs of genes/proteins; ii) the dimension of the gene arrangements (genes on a list or arranged on a plane); iii) the number of isothermal steps for the annealing procedure in gene ordering; iv) the number of temperature changes (each change represents lowering the temperature value to 50% of its current value); v) printing interval is the number of Monte Carlo steps between printing steps; vi) the value of parameter α in Eq. (1).

The outputs are the snapshots of the association matrix, for 1d only, a plain text two columns files containing the positions of each interacting protein pair, and the ordered proteins list, a plain text two (1d) or three (2d) columns file. The name of those output files are, respectively, ”associationmatrix_” and “ordering_” followed by the input file name .and it is written at the same directory as the input fiile.

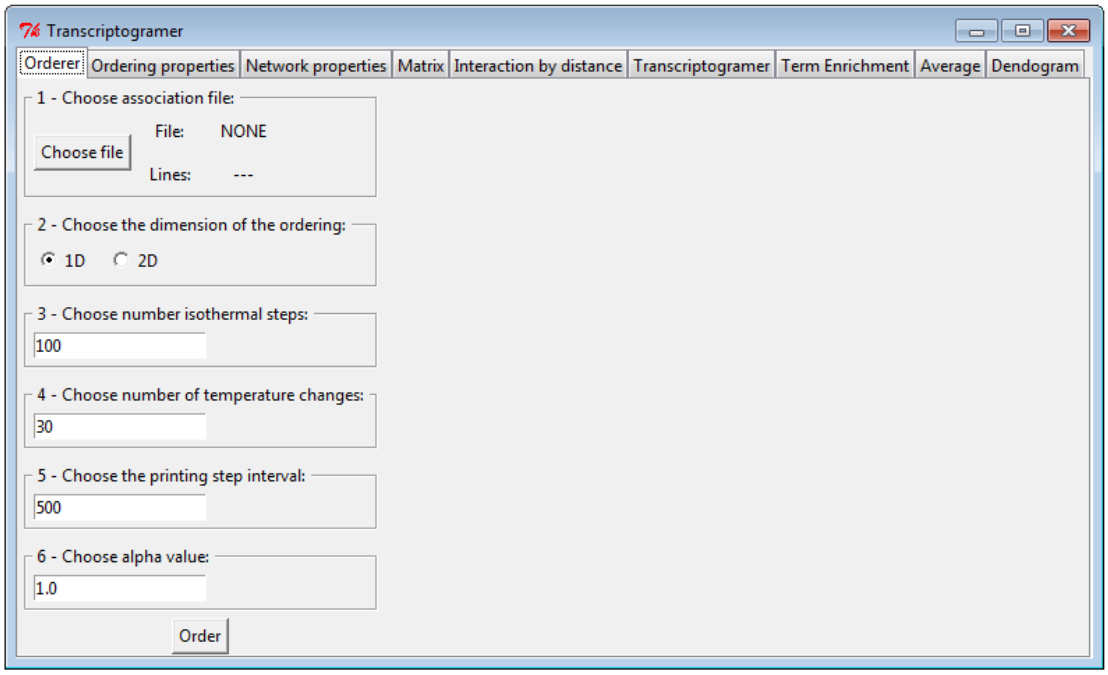


Figure 1a. The Transcriptogramer first tab: The orderer.

ENSP00000000412	ENSP00000221957
ENSP00000000442	ENSP00000206249
ENSP00000000442	ENSP00000229022
ENSP00000000442	ENSP00000233557
ENSP00000000442	ENSP00000243050
ENSP00000000442	ENSP00000246672
ENSP00000000442	ENSP00000253727
ENSP00000000442	ENSP00000254066
.....	.....

Figura 1b. Association file, the input file for Tab1: plain text, two column file, no headings, with rows informing the pairs of genes/proteins that are associated. The gene/protein identification used in this file defines gene/protein identification for all other Tabs.

Protein	dim1
ENSP00000361562	1
ENSP00000345728	2
ENSP00000053867	3
ENSP00000316854	4
ENSP00000261249	5
ENSP00000242839	6
ENSP00000017003	7
ENSP00000029410	8
ENSP00000230053	9
ENSP00000265471	10

Figura 1c. Ordered gene/protein list, the output file for Tab1, option one dimension: plain text, two columns file, with a heading line, with rows informing the gene/protein identification and its position on the ordered gene/protein list. The gene/protein identification used in this file is the same as used in the input file.

Protein	dim1	dim2
ENSP00000000412	76	141
ENSP00000000442	95	80
ENSP00000001008	80	127
ENSP00000001146	125	68
ENSP00000002596	73	143
ENSP00000002829	100	145
ENSP00000003084	84	115
ENSP00000003100	114	54

Figura 1d. Ordered gene/protein list, the output file for Tab1, option two dimensions: plain text, three columns file, with a heading line, with rows informing the gene/protein identification and its position on the planar gene/protein arrangement The gene/protein identification used in this file is the same as used in the input file.

TAB 2 - Ordering properties

This step produces profiles of graph properties projected on the ordered arrangement of genes/proteins.

The arrangement of genes by the probability that they are associated builds up regions in the gene list or planar arrangement that have different characteristics. To evaluate these ordering-dependent features, The Transcriptogramer calculates in TAB 2 - Ordering properties three quantities: window modularity, connectivity, and clustering coefficient, defined as follows:

2.1 Window modularity.

Modularity of a subset of nodes in a graph is defined as the ratio of the number of links between the edges linking any two nodes in the subset by the number of links that connects at least one of the nodes of the subset. Modularity is a number in the interval [0,1], assuming 0 if the nodes in the subset do not present any link, and assuming 1 when all links presented by the nodes in the subset are with another node of the subset.

Window modularity is a property assigned to all genes/proteins of the ordered list or planar arrangement and is defined as the modularity of the subset defined by the gene/protein and its neighbors inside a region of radius r centered at the gene/protein.

2.2 Connectivity

Connectivity or degree of a node is the number of links it presents. The connectivity assigned to each gene here is the average connectivity calculated over all genes/proteins a inside a region of radius r centered at the gene/protein.

2.3 Clustering coefficient

Clustering coefficient of a node of degree k is the ratio between the number existing links between the k neighbors of the gene/proteins and the maximum possible number of such links , $\frac{k(k-1)}{2}$. The Clustering coefficient assigned to each gene here is the average clustering coefficient calculated over all genes/proteins inside a inside a region of radius r centered at the gene/protein.
The input files for this tab are the association and the ordering files, as described in the previous section and illustrated in Figs. 1b-d.

The output file is a plain text, a five (1d) or six columns (2d) file, with a heading, with the rows containing the gene/protein identification, its position on the list or planar arrangement, and the assigned values of window modularity, average connectivity and clustering coefficients. See Fig. 2.b

The name of the output is "orderingProperties" followed by the input transcriptome file name .and it is written at the same directory as the input fiile.

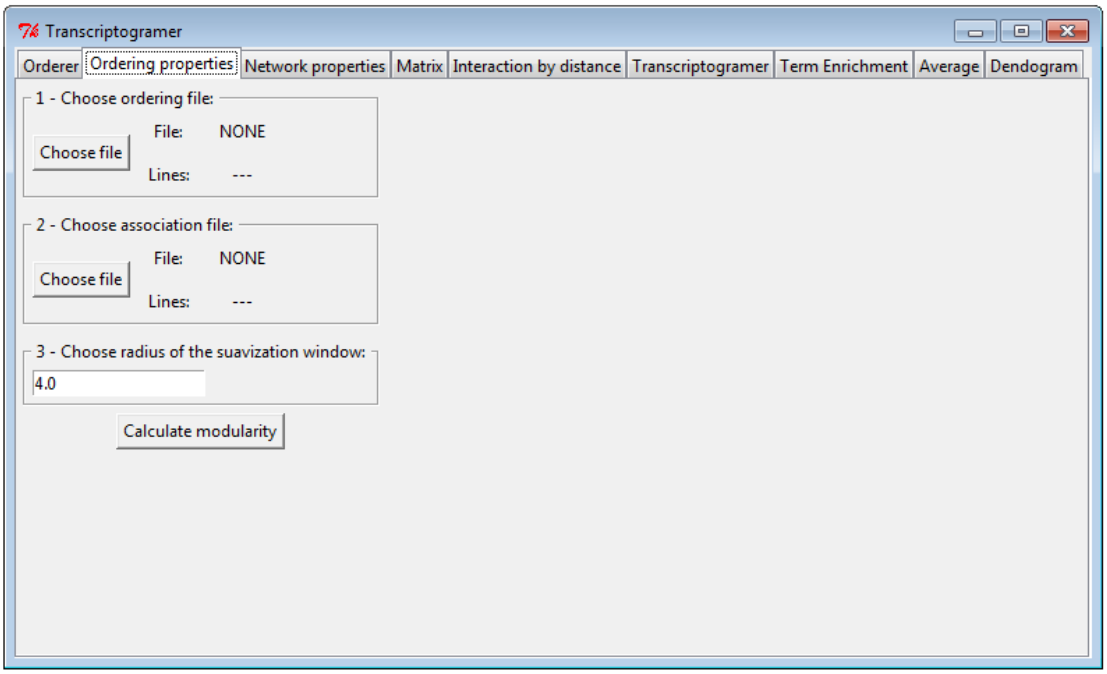


Figure 2a. The Transcriptogramer second tab: Ordering properties

#radius = 50				
Protein	position(dim1)	modularity	connectivity	clustering
ensp00000306185	1	0.987654	1.58824	0.222222
ensp00000317618	2	0.987952	1.59615	0.217949
ensp00000362171	3	1	1.58491	0.213836
ensp00000301945	4	0.965517	1.61111	0.228395
ensp00000331106	5	0.955556	1.63636	0.242424
ensp00000362616	6	0.945055	1.625	0.238095
ensp00000268124	7	0.934783	1.61404	0.233918
ensp00000303511	8	0.924731	1.60345	0.229885

Figura 2b. Ordering properties, the output file for TAB2, option one dimension: plain text, five columns file, with two heading lines, the first indicates the radius or the region used for averaging, with rows informing the gene/protein identification, its position on the gene/protein list and the assigned values for window modularity and average connectivity and clustering coefficient. The gene/protein identification used in this file is the same as used in the association file. Two dimension option would contain an extra position column.

TAB 3 - Network properties

This step produces calculates average graph properties as a function of the node connectivity.

The Transcriptogramer offers the possibility of calculating network properties as a function of the gene/protein connectivity (Fig.3a). TAB3 - Network properties input is the association file, as discussed in the section TAB1 - The ordered and illustrated in Fig.1b. This file lists the pairs of genes/proteins that are considered as associated.

The calculated network properties are

4.1 Probability that a gene/protein $P(k)$ has connectivity k.

4.2 Assortativity of a graphs is measured by the average connectivity of the neighbors of a node with connectivity k.

4.3 Clustering coefficient of a node of degree k is the ratio between the number existing links between the k neighbors of the gene/proteins and the maximum possible number of such links , $\frac{k(k-1)}{2}$. The Transcriptogramer calculates in TAB3 the average clustering coefficient of the nodes of degree k.

The output is a plain text, four columns file, with a heading line, as illustrated in Fig. 3b.

The name of the output is "networkProperties_" followed by the input association file name and it is written at the same directory as the input fiile.

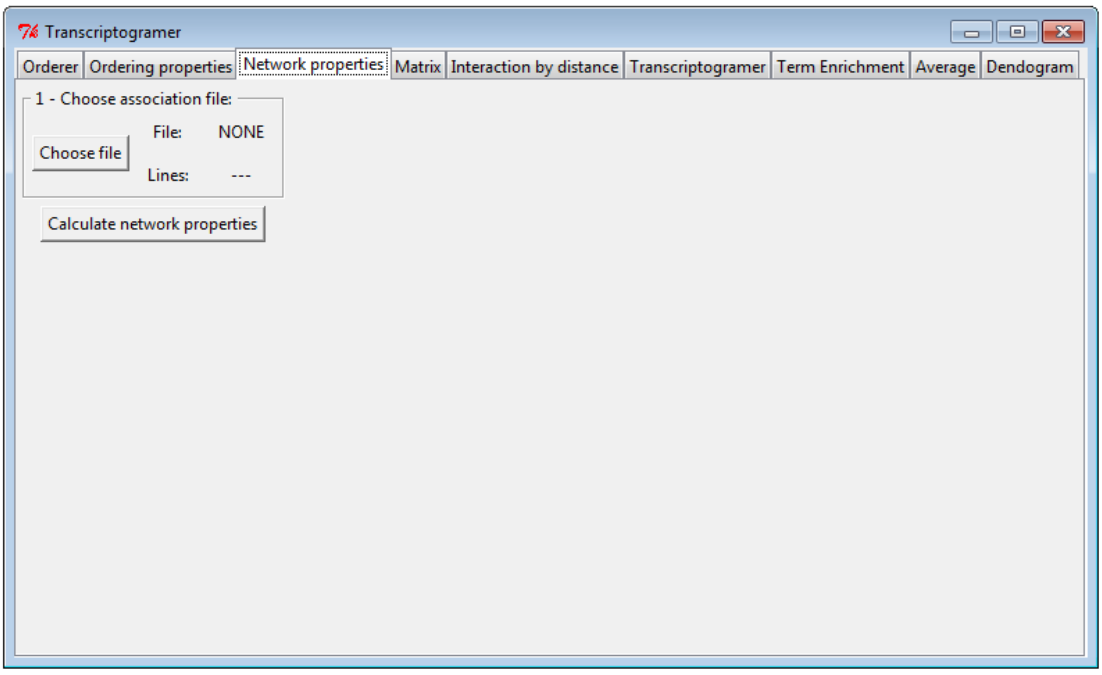


Figure 3a. The Transcriptogramer third tab: Network properties.

k	p(k)	assortativity(k)	clustering(k)
1	0.139421	409.723	0
2	0.0807714	442.128	0.518258
3	0.0680658	912.256	0.448333
4	0.0441293	385.578	0.526564
5	0.0431083	358.621	0.585789
6	0.0380034	328.547	0.58806
7	0.0271129	37.792	0.52899

Figura 3b. Network properties, the output file for TAB3: plain text, four columns file, with a heading line, with rows informing the gene/protein identification, the value of connectivity k, the probability of finding a gene/protein, the average connectivity of the neighbors and the average clustering coefficients for genes/proteins with connectivity k.

TAB 4 - Matrix

This step produces a file containing the position on the one dimensional ordered list of the pairs of associated genes/proteins. This file may be used to make an image of the association matrix.

The inputs for this function are the ordering and the association file, as described in the section for TAB1 -The orderer. The idea is to provide the plotting of the association matrix, which gives an image of modularity and clustering properties of the ordered gene list. This option maybe applied for one dimension ordering list only.

The output is plain text, two columns, no headings, each row informing the position on the list of the associated pair of genes/proteins.

The name of the output is “associationmatrix_” followed by the input file name and it iswritten at the same directory as the input fiile.

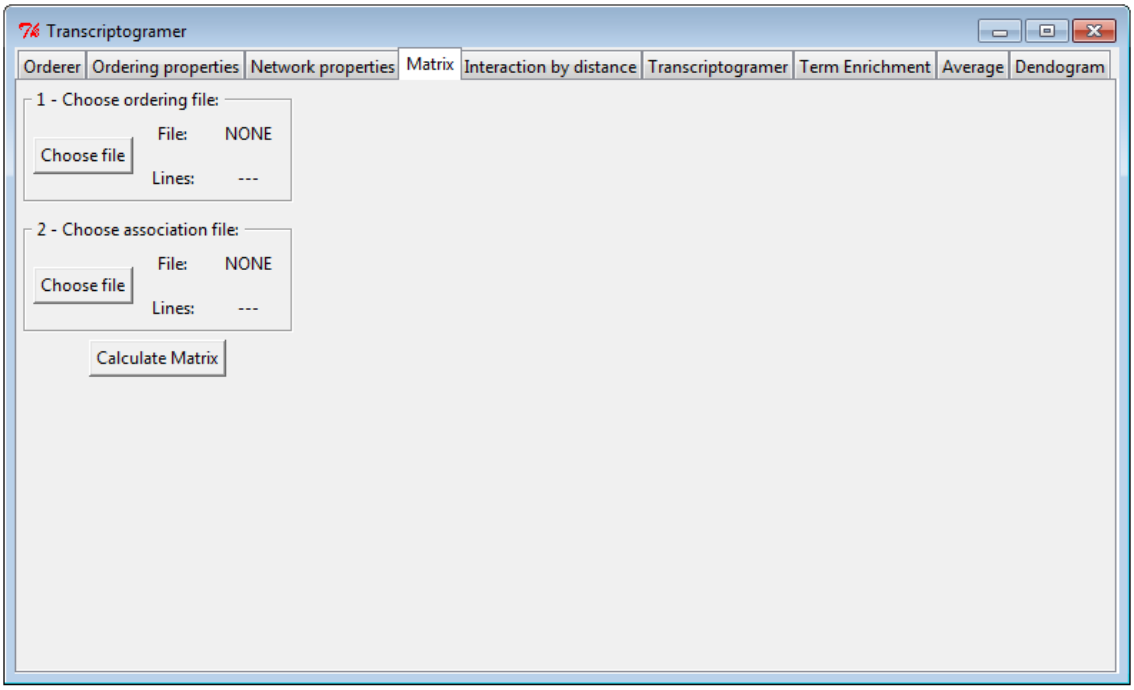


Figure 4. The Transcriptogramer forth tab: Matrix. It applies for one dimensional arrangements only.

TAB 5 - Interactions by distance

This step calculates the fraction of associated genes/proteins that are at a given distance on the ordered arrangement.

The inputs for this function are the ordering and the association file, as described in the section for TAB1 -The orderer.

The output is plain text, two columns, one heading line, each row informing a distance and the average number of associated genes that are at that distance on the ordered list or planar arrangement.

The name of the output is “interactions_” followed by the input file name and it is written at the same directory as the input fiile.

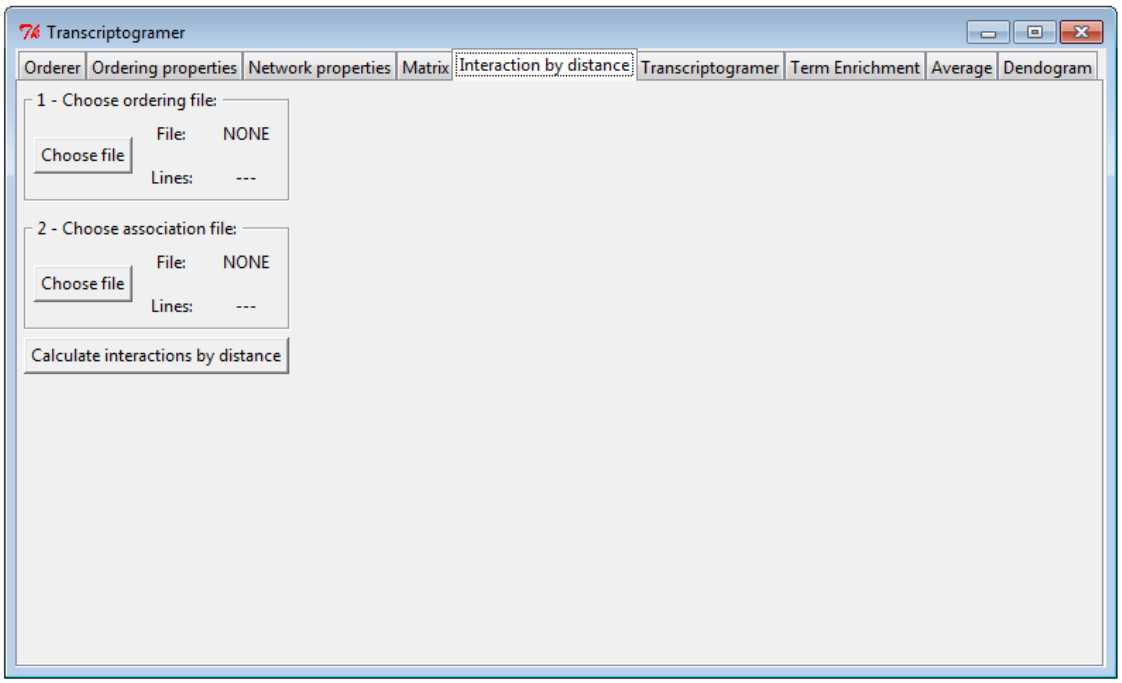


Figure 5. The Transcriptogramer fifth tab: Interactions by distance.

TAB 6 - Transcriptogramer

This step produces transcriptograms.

This TAB6 produces transcriptograms, that is, it projects gene expression data on the ordered genes/proteins list or planar arrangement. For each transcriptome sample, the transcriptogramer follows the steps as described below.

i) Assign to each gene/protein the expression data, averaging over all probes sets related to the same gene/protein. For this step it is necessary to feed the program with a dictionary that links probe sets to genes/proteins;

ii) To each position of the ordered gene list or planar arrangement, assign a value equal to the average of the expression levels inside the region of radius r around that position.

The input may contain information on one or more samples; The Transcriptogramer automatically calculates the transcriptogramer for all samples in the expression file.

The inputs are the ordering file as produced by The orderer TAB, a dictionary that links the probsets of the microarray platform used for producing the expression data, and the value of the desired radius of the region over which averages of expression levels should be performed. Figure 6a presents the appearance of TAB 6 - Transcriptogramer.

The dictionary file is a plain text file, with a heading line . The first column specifies the probeset identification and the following columns specify the expression levels relative to different samples. See Figure 6b for an example.

The expression file is a plain text, two columns file, with no headings . The first column contains the gene/protein identification while the second one lists the associated probe sets identifications. See Fig. 6c for an example.

The output is a plain text file, two heading lines as shown in Fig.6d. The first colum provides genes/proteins identification, as provided by the ordering file. The second column provides the position of the gene on the ordered list for the one dimensional option. For the two dimensional option, there are two position columns. The other columns are the transcriptograms relative to the samples as provided by the expression file. See Figs. 6c and 6d.

The name of the output is “transcriptogram_” followed by the input transcriptome file name and it is written at the same directory as the input transcriptome fiile.

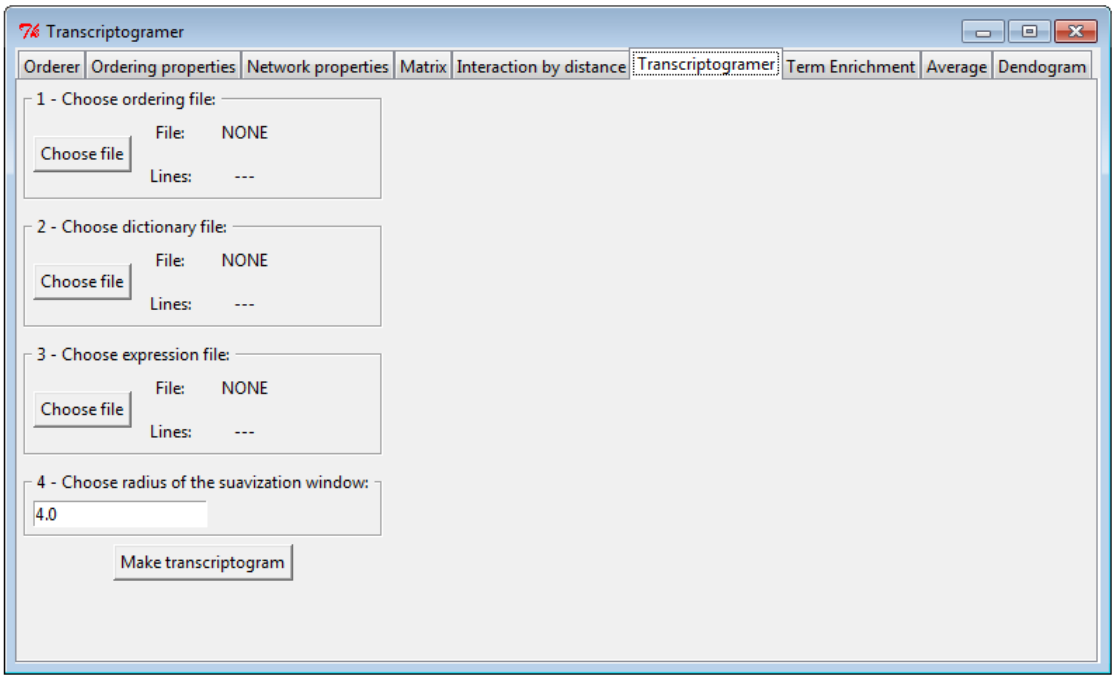


Figure 6a. The Transcriptogramer sixth tab: Transcriptogramer.

ENSP00000392314	202587_s_at
ENSP00000448741	202587_s_at
ENSP00000449130	202587_s_at
ENSP00000362271	202588_at
ENSP00000314727	202589_at
ENSP00000314902	202589_at
ENSP00000315644	202589_at
ENSP00000007708	202590_s_at
ENSP00000420927	202590_s_at

Figura 6b. Input: Dictionary file. Plain text, two columns, no headings, first column with gene/protein identification (the same identification as in the ordering file).

!Sample_title	Normal-1	Normal-2	Normal-3	LS-1	LS-2	LS-3
1007_s_at	8.927225959	9.323152502	9.425702496	9.132472598	8.821365725	8.4035751
1053_at	6.806054169	6.130473175	5.169634534	7.337749272	7.46242396	6.703927603
117_at	5.843561866	5.842975454	4.749799793	5.324129244	5.524537584	5.254838691
121_at	5.557897749	5.171537563	5.547484327	5.662241225	6.244414944	6.311535284
1255_g_at	2.104409419	2.081901801	2.09300007	2.108852437	2.092603764	2.147344546
1294_at	4.949555182	4.094465109	4.30166568	5.018234151	5.18919056	4.528377161
1316_at	4.049359939	5.513031349	4.145075272	3.572123882	3.64451964	3.728640093
1320_at	5.828528862	4.760341329	4.312674068	3.258123575	4.661248199	3.150818043

Figure 6c. Input: Expression file. Plain text, one heading line. First column for probe set identification and as many columns as the number of samples in the expression dataset.

#radius = 4							
Protein	Position(dim1)	Normal-1	Normal-2	Normal-3	LS-1	LS-2	LS-3
ensp00000361562	1	8.76451	8.24171	8.03078	9.03263	9.19896	8.83332
ensp00000345728	2	7.73456	7.19186	7.01791	7.8735	7.99172	7.6896
ensp00000053867	3	7.34759	6.95682	6.69537	7.39059	7.48311	7.27758
ensp00000316854	4	7.09426	6.76884	6.51669	7.14257	7.18807	7.04189
ensp00000261249	5	6.78986	6.35183	6.21728	6.77384	6.79288	6.74911
ensp00000242839	6	6.26074	5.98081	5.67684	6.13827	6.09359	6.09631
ensp00000017003	7	6.75593	6.5661	6.20219	6.76128	6.65457	6.63906

Figure 6d. Output: The transcriptograms for one dimensional ordering. Plain text, many columns file. First heading indicates the radius or the region used for averaging the expression levels. Second column informs the location of the gene/protein on the ordering. The other columns are the transcriptograms for each sample in the input expression file.

TAB 7 - Term enrichment

This function calculates the occupation density of a gene/protein set.

Given a set of genes, TAB 7 calculates the fraction of sites is occupied by the genes/proteins in the set inside a region of radius r around a central position. The occupation fraction value is assigned to the central position. This function is useful to localize terms of Gene Ontology or metabolic pathways on the gene/protein arrangement.

The files containing the list of genes/proteins should be plain text, no heading, containing the genes/proteins identification to be localized in the third column, which is the "gene association format", the output format option from Gene Ontology. One file for each different gene/protein list.

The inputs for the function Term enrichment are

i) the ordering file, as described in the section relative to TAB1 - The orderer;

ii) a dictionary file that links the gene/protein identification used in the ordering file with the identification used in the lists of genes/proteins to be localized. In case they are the same, the dictionary is a plain text file containing two identical columns ;

iii) a file list, which is a plain text, one column file, no heading, containing the name of the files with the gene/protein lists to be localized. Observe that these files must be previously produced;

iv) the value of radius r of the region inside which the occupation density is calculated.

.

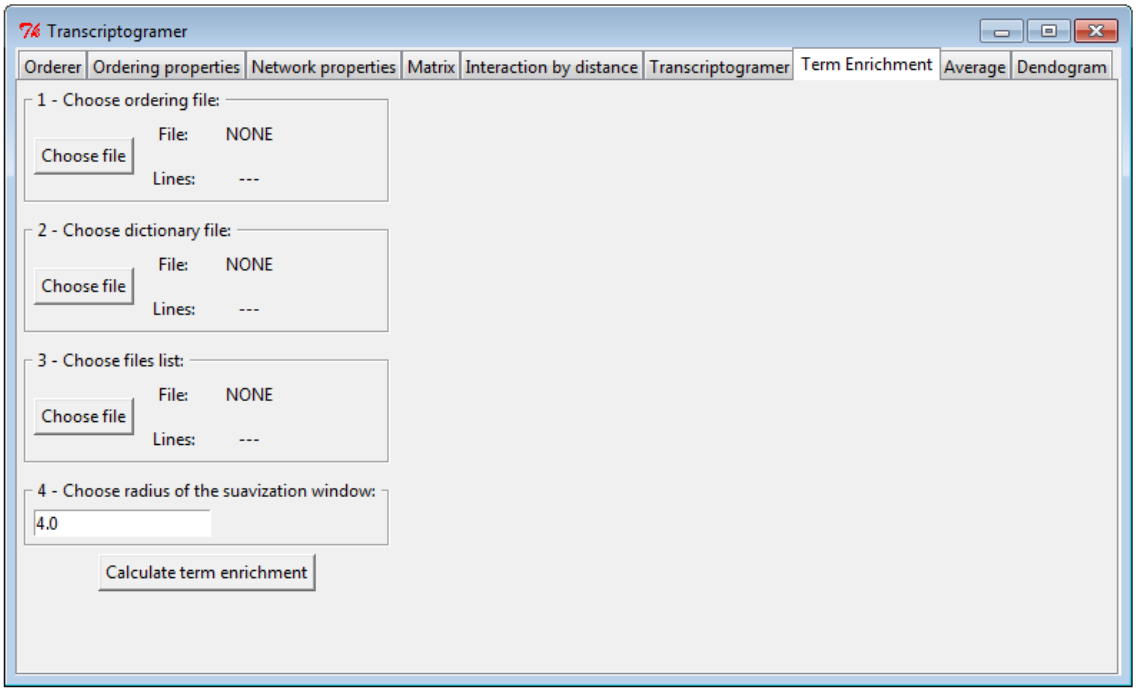


Figure 7a. The Transcriptogramer seventh tab: Term enrichment.

ENSP00000306625	ASPSCR1
ENSP00000344020	ASPSCR1
ENSP00000301776	ASRGL1
ENSP00000400057	ASRGL1
ENSP00000431772	ASRGL1
ENSP00000433136	ASRGL1
ENSP00000443284	ASRGL1
ENSP00000253004	ASS1

Figure 7b. The input file dictionary. Plain text, two columns, no heading. First column gene/protein identification used in ordering file. Second column, gene/protein identification used in the lists of genes to be localized (GO terms, metabolic pathways, customized list).

TAB 8 - Averages

This function calculates the average transcriptogram and the standard deviation of a group of transcriptograms.

To define a group of samples, it is sufficient to stipulate a common character string in the headings of the group sample that does not appear in any other sample. Figure 9.a shows the TAB 9 - Average whith the specification of two groups of samples: the first one contains 'normal' in the samples headings and the second group contains the string 'LS'.

The input file is the transcriptogram file produce by TAB6 - Transcriptogramer.

The output file is a plain text, multicolumn file, with one heading line. The first column contains the gene/protein id as in the ordering file, the second the gene position for the one dimension option or the second and third columns contain the gene/protein position on the planar arrangement. The following columns contain the average and the standard deviations for each group.

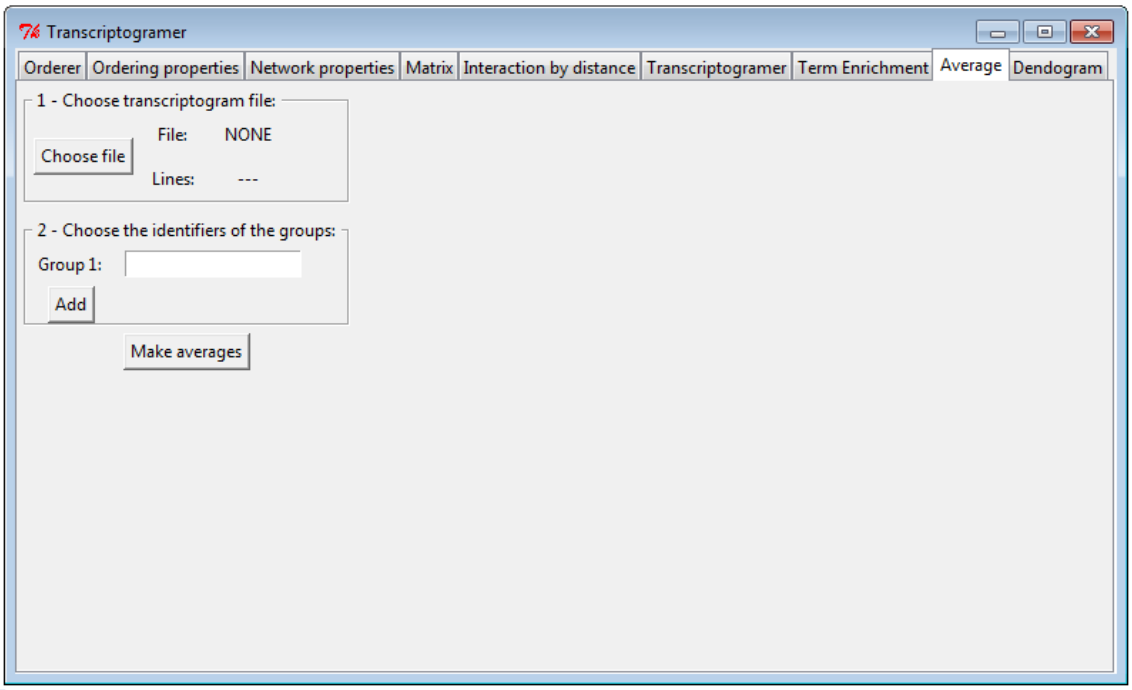


Figure 8a. Tab 8 - average where two groups have been defined to calculate the average and the standard deviation. The group identification is some set of characters that are on the heading of all sample transcriptogram and, at the same time, is absent of any other column heading.

References

[daSilva2013a]da Silva, S.R, Perrone, G.C., de Almeida R.MC. Transcriptograms strategies and statistics. To appear.(2013)

[daSilva2013b] da Silva, S.R, Perrone, G.C., de Almeida R.MC. The Transcriptogramer. To appear. (2013)

[Edgar2002] Edgar, R; Domrachev, M; Lash, AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30,** 207-210 DOI: 10.1093/nar/30.1.207 (2002).

[GeneOntology2000] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat. Genet.;25(1):25-9. (2000).

[Jensen2009] Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412-D416. (2009)

[Metropolis1949] Metropolis, N. and Ulam, S. The Monte Carlo Method . *J. Amer. Stat. Assoc.* **44** (247): 335–341. (1949).

[Kirkpatrick1983] Kirkpatrick, S.; C. D. Gelatt, M. P. Vecchi. Optimization by Simulated Annealing. *Science*. New Series **220** (4598): 671–680. (1983)

[Perrone2013] Perrone, G.C., da Silva, S.R., de Almeida, R.M.C. Transcriptograms in two dimensions. To appear.(2013).

[Rybarczyk2011] Rybarczyk-Filho, J.L., Castro, M.A.A., Dalmolin, R.J, Moreira, J.C.F., Brunnet, L.G. and de Almeida, R.M.C., Towards a genome-wide transcriptogram: the *Saccharomyces cerevisiae* case. *Nucleic Acids Res.*, 39, 3005-3016 (2011).PMID:21169199

Bibliografia

[1] Rybarczyk-Filho, J. L., Castro, M. A. A., Dalmolin, R. J. S., Moreira, J. C. F., Brunnet, L. G., and de Almeida, R. M. C. Towards a genome-wide transcriptogram: the Saccharomyces cerevisiae case. *Nucleic Acids Res.* **39**(8), 3005–3016, APR (2011).

[2] Freeman, W., Robertson, D., and Vrana, K. Fundamentals of DNA hybridization arrays for gene expression analysis. *Biotechniques* **29**(5), 1042+, NOV (2000).

[3] Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**(13), 1675–1680, DEC (1996).

[4] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**(SI), D277–D280, JAN 1 (2004).

[5] Nair, R. P., Duffin, K. C., Helms, C., Ding, J., Stuart, P. E., Goldgar, D., Gudjonsson, J. E., Li, Y., Tejasvi, T., Feng, B.-J., Ruether, A., Schreiber, S., Weichenthal, M., Gladman, D., Rahman, P., Schrodi, S. J., Prahalad, S., Guthery, S. L., Fischer, J., Liao, W., Kwok, P.-Y., Menter, A., Lathrop, G. M., Awise, C., Begovich, A. B., Voorhees, J. J., Elder, J. T., Krueger, G. G., Bowcock, A. M., Abecasis, G. R., and Psoriasi, C. A. S. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappa B pathways. *Nat. Genet.* **41**(2), 199–204, FEB (2009).

[6] Swindell, W. R., Johnston, A., Carbajal, S., Han, G., Wohn, C., Lu, J., Xing, X., Nair, R. P., Voorhees, J. J., Elder, J. T., Wang, X.-J., Sano, S., Prens, E. P., DiGiovanni, J., Pittelkow, M. R., Ward, N. L., and Gudjonsson, J. E. Genome-Wide Expression Profiling of Five Mouse Models Identifies Similarities and Differences with Human Psoriasis. *Plos One* **6**(4), APR 4 (2011).

[7] Watson, J. and Crick, F. Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid. *Nature* **171**(4356), 737–738 (1953).

[8] Wang, Z., Gerstein, M., and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**(1), 57–63, JAN (2009).

[9] Heid, C., Stevens, J., Livak, K., and Williams, P. Real time quantitative PCR. *Genome Res.* **6**(10), 986–994, OCT (1996).

[10] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**(D1), D808–D815, JAN (2013).

[11] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**(1), D561–D568, JAN (2011).

[12] von Mering, C., Jensen, L., Snel, B., Hooper, S., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M., and Bork, P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**(SI), D433–D437, JAN 1 (2005).

[13] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G., and Consortium, G. O. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**(1), 25–29, MAY (2000).

[14] Edgar, R., Domrachev, M., and Lash, A. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**(1), 207–210, JAN 1 (2002).

[15] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.* **41**(D1), D991–D995, JAN (2013).

[16] Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., Kurbatova, N., Malone, J., Mani, R., Mupo, A., Pereira, R. P., Pilicheva, E., Rung, J., Sharma, A., Tang, Y. A., Ternent, T., Tikhonov, A., Welter, D., Williams, E., Brazma, A., Parkinson, H., and Sarkans, U. ArrayExpress update-trends in database growth and links to data analysis tools. *Nucleic Acids Res.* **41**(D1), D987–D990, JAN (2013).

[17] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2), 249–264, APR (2003).

[18] Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bionformatics* **19**(2), 185–193, JAN 22 (2003).

[19] Lim, W. K., Wang, K., Lefebvre, C., and Califano, A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bionformatics* **23**(13), I282–I288, JUL 1 (2007). 15th Conference on Intelligent Systems for Molecular Biology/6th European Conference on Computational Biology, Vienna, AUSTRIA, JUL 21-25, 2007.

[20] Gyorffy, B., Molnar, B., Lage, H., Szallasi, Z., and Eklund, A. C. Evaluation of Microarray Preprocessing Algorithms Based on Concordance with RT-PCR in Clinical Samples. *Plos One* **4**(5), MAY 21 (2009).

[21] Pepper, S. D., Saunders, E. K., Edwards, L. E., Wilson, C. L., and Miller, C. J. The utility of MAS5 expression summary and detection call algorithms. *BMC Bionformatics* **8**, JUL 30 (2007).

[22] Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., and Spencer, F. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99**(468), 909–917, DEC (2004).

[23] Wu, C., Zhao, H., Baggerly, K., Carta, R., and Zhang, L. Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. *Bionformatics* **23**(19), 2566–2572, OCT 1 (2007).

[24] Gautier, L., Cope, L., Bolstad, B., and Irizarry, R. affy - analysis of Affymetrix GeneChip data at the probe level. *Bionformatics* **20**(3), 307–315, FEB 12 (2004).

[25] Gentleman, R. C., Carey, V. J., Bates, D. M., and others. Bioconductor: Open software development for computational biology and bionformatics. *Genome Biol.* **5**, R80 (2004).

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2013). ISBN 3-900051-07-0.

[27] METROPOLIS, N. and ULAM, S. THE MONTE CARLO METHOD. *J. Am. Stat. Assoc.* **44**(247), 335–341 (1949).

[28] Rybarczyk-Filho, J. L. *Medidas de performance metabólica usando a expressão gênica de genoma completo*. PhD thesis, UFRGS, (2011).

[29] Perrone, G. C. Transcriptograma em duas dimensões. Master's thesis, UFRGS, (2013).

[30] Vinogradov, A. E. Modularity of cellular networks shows general center-periphery polarization. *Bioinformatics* **24**(24), 2814–2817, DEC 15 (2008).

[31] Balanda, K. and Macgillivray, H. Kurtosis - a critical review. *Am. Stat.* **42**(2), 111–119, MAY (1988).

[32] Hwang, S., Kwak, S. H., Bhak, J., Kang, H. S., Lee, Y. R., Koo, B. K., Park, K. S., Lee, H. K., and Cho, Y. M. Gene Expression Pattern in Transmitochondrial Cytoplasmic Hybrid Cells Harboring Type 2 Diabetes-Associated Mitochondrial DNA Haplogroups. *Plos One* **6**(7), JUL 13 (2011).

[33] Shi, L., Reid, L. H., and Jones, W. D. e. a. *Nat. Biotechnol.* **24**(9), 1151–1161, SEP (2006).

[34] McDonald, J. H. *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, (2009).

[35] Shi, L., Perkins, R. G., Fang, H., and Tong, W. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr. Opin. Biotechnol.* **19**(1), 10–18, FEB (2008).

[36] Shi, L., Jones, W. D., Jensen, R. V., Harris, S. C., Perkins, R. G., Goodsaid, F. M., Guo, L., Croner, L. J., Boysen, C., Fang, H., Qian, F., Amur, S., Bao, W., Barbacioru, C. C., Bertholet, V., Cao, X. M., Chu, T.-M., Collins, P. J., Fan, X., Frueh, F. W., Fuscoe, J. C., Guo, X., Han, J., Herman, D., Hong, H., Kawasaki, E. S.,

Li, Q.-Z., Luo, Y., Ma, Y., Mei, N., Peterson, R. L., Puri, R. K., Shippy, R., Su, Z., Sun, Y. A., Sun, H., Thorn, B., Turpaz, Y., Wang, C., Wang, S. J., Warrington, J. A., Willey, J. C., Wu, J., Xie, Q., Zhang, L., Zhang, L., Zhong, S., Wolfinger, R. D., and Tong, W. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bionformatics* **9**(9) (2008). 5th Annual Conference of the MidSouth-Computational-Biology-and-Bionformatics-Society, Oklahoma City, OK, FEB 23-24, 2008.

[37] Lin, G., He, X., Ji, H., Shi, L., Davis, R. W., and Zhong, S. Reproducibility Probability Score - incorporating measurement variability across laboratories for gene selection. *Nat. Biotechnol.* **24**(12), 1476–1477, DEC (2006).

[38] Lin, L. A concordance correlation-coefficient to evaluate reproducibility. *Biometrics* **45**(1), 255–268, MAR (1989).

[39] Popovici, V., Chen, W., Gallas, B. G., Hatzis, C., Shi, W., Samuelson, F. W., Nikolsky, Y., Tsyganova, M., Ishkin, A., Nikolskaya, T., Hess, K. R., Valero, V., Booser, D., Delorenzi, M., Hortobagyi, G. N., Shi, L., Symmans, W. F., and Pusztai, L. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.* **12**(1) (2010).

[40] Shi, L., Campbell, G., and Jones, W. D. e. a. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Pharmacogenomics J.* **8**(S), S5–S16, OCT (2010).

[41] Meyer, P., Hoeng, J., Rice, J. J., Norel, R., Sprengel, J., Stolle, K., Bonk, T., Corthesy, S., Royyuru, A., Peitsch, M. C., and Stolovitzky, G. Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. *Bionformatics* **28**(9), 1193–1201, MAY 1 (2012).

[42] Meyer, P., Alexopoulos, L. G., Bonk, T., Califano, A., Cho, C. R., de la Fuente, A., de Graaf, D., Hartemink, A. J., Hoeng, J., Ivanov, N. V., Koeppl, H., Linding, R., Marbach, D., Norel, R., Peitsch, M. C., Rice, J. J., Royyuru, A., Schacherer, F., Sprengel, J., Stolle, K., Vitkup, D., and Stolovitzky, G. Verification of systems biology research in the age of collaborative competition. *Nat. Biotechnol.* **29**(9), 811–815, SEP (2011).

[43] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, second edition edition, (2010).

[44] Carey, V., Gentleman, R., Mar, J., , contributions from Jason Vertrees, and Gatto, L. *MLInterfaces: Uniform interfaces to R machine learning procedures for data in Bioconductor containers*. R package version 1.34.2.

[45] Yao, Y., Richman, L., Morehouse, C., de los Reyes, M., Higgs, B. W., Boutrin, A., White, B., Coyle, A., Krueger, J., Kiener, P. A., and Jallal, B. Type I Interferon: Potential Therapeutic Target for Psoriasis? *Plos One* **3**(7), JUL 16 (2008).

[46] Brynedal, B., Khademi, M., Wallström, E., Hillert, J., Olsson, T., and Duvefelt, K. Gene expression profiling in multiple sclerosis: A disease of the central nervous system, but with relapses triggered in the periphery? . *Neurobiol. Dis.* **37**(3), 613 – 621 (2010).

[47] Benetti, F. P. d. C. Homo sapiens: análise de expressão gênica por transcriptograma (Trabalho de conclusão de curso), (2010).

[48] Bigler, J., Rand, H. A., Kerkof, K., Timour, M., and Russell, C. B. Cross-study homogeneity of psoriasis gene expression in skin across a large expression range. *Plos One* **8**(1), e52242, 01 (2013).

[49] Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L. J., and Bork, P. STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Res.* **40**(D1), D876–D880, JAN (2012).