Universidade Federal do Rio Grande do Sul
Instituto de Física

Transcriptograma em duas dimensões

Gabriel Cury Perrone

Dissertação de mestrado

Porto Alegre, Fevereiro de 2013

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

Transcriptograma em duas dimensões[†]

Gabriel Cury Perrone

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

Porto Alegre, Fevereiro 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)

Agradecimentos

- Aos Professores Rita M. C. de Almeida e Leonardo G. Brunnet pela amizade, apoio e dedicação ao me orientar.
- Ao meu irmão e amigo, que me incentivou a estudar e a buscar novas soluções para os problemas que surgiram ao longo deste projeto. Nossas discussões trouxeram ideias e abordagens diversas.
- Aos meus pais pelo apoio e carinho incondicional.
- À minha família, que me acompanhou durante todo o curso com muito carinho.
- Aos meus colegas e amigos da UFRGS, minha família dentro da universidade.
- Aos meus amigos e Fellow Geeks, minha família fora da universidade, me mantendo louco o suficiente para seguir este caminho.
- Enfim, agradeço a todos os que contribuíram para a construção do caminho que tracei.

Resumo

O conhecimento sobre o Genoma está crescendo rapidamente, assim como a quantidade de técnicas de medida da expressão gênica. É sabido que decodificar o DNA não é suficiente para entender o metabolismo celular e suas alterações, para isso, precisamos entender a expressão dos genes. Existem técnicas de medida de expressão de genoma completo, mas estas possuem ruído muito elevado, dificultando a análise dos seus resultados. Devido a isto, foram desenvolvidas técnicas para analisar estes resultados e aumentar a razão sinal-ruído; entre estas, temos o Transcriptograma.

O Transcriptograma é dividido em duas etapas: o ordenamento de redes e o cálculo de médias. O ordenamento é feito por minimização de função custo, trata o Proteoma como uma rede simples e a ordena em uma lista utilizando o método de Monte Carlo para aproximar proteínas associadas. A partir da rede ordenada é possível analisar as propriedades desta, como a sua distribuição de módulos e de processos biológicos, e calcular o Transcriptograma através do cálculo das médias dos valores de expressão das proteínas dentro de uma vizinhança. Este método reduz o ruído das medidas de expressão gênica e possibilita a análise de performance celular, descrevendo o estado das células no momento da medida.

Nesta dissertação, aprimoramos o método original de ordenamento alterando profundamente seu algoritmo. As modificações efetuadas reduziram em mais de mil vezes o tempo de execução do programa. As alterações obtidas abriram a possibilidade de ordenar a rede em uma dimensão qualquer, então produzimos um novo programa para obter o ordenamento da rede de proteínas em duas dimensões. Analisamos os novos resultados observando a distribuição dos módulos e de funções biológicas. A generalização do transcriptograma para duas dimensões mostra resultados melhores que os obtidos a partir do ordenamento em uma lista.

Também propomos um método de seleção de amostras em duas classes e o aplicamos ao diagnóstico de Psoríase. Esse método separou claramente as amostras saudáveis das doentes. Com a rede ordenada,é possível analisar as regiões que mostram alterações no Transcriptograma e observar quais funções biológicas estão alteradas, obtendo mais informações sobre o estado celular e possibilitando a descoberta de novos alvos para fármacos.

Abstract

Knowledge about the Genome is growing rapidly, as well as the number of techniques for measuring gene expression. It is known that decoding the DNA is not sufficient to understand cell metabolism and its alterations, for that we need to understand the expression of the genes. There are techniques for measuring expression of the complete genome, but these have very high noise, making it hard to analyze the results. Because of that, techniques were developed to analyze these results and increase the signal to noise ratio. Among these techniques there is the Transcriptogram.

The Transcriptogram is divided into two stages: the ordering of networks and the calculation of the average of Transcriptomes. The ordering is made by minimizing a cost function, it treats the Proteome as a simple network and orders it in a list using the Monte Carlo method to make closer proteins that are associated. From the ordered network it is possible to analyze its properties, such as its modules and biological processes distribution, and calculate the Transcriptogram by calculating the mean value of protein expression in a neighborhood. This method reduces the noise of the measurements of gene expression and enables the analysis of cell performance, describing the state of the cells at the time of measurement.

In this dissertation we improved the original ordering method making deep changes in its algorithm. These changes reduced the program execution time in more than one thousand times. These alterations openned the possibility of ordering the network in any dimension, then we produced a new programm to obtain the network ordering in two dimensions. We analyzed the new results by observing the modules and biological functions distribution. The generalization of the Transcriptogram to two dimensions shows better results than those obtained from the ordering in a list.

We also propose a method for selecting samples into two classes and apply it to the diagnosis of Psoriasis. It clearly separated the samples from healthy patients. With the network ordered, we can analyze the regions that show alterations in the Transcriptogram and observe which biological functions are altered, obtaining more information about cell state and enabling the discovery of new targets for drugs.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Em 1953, foi desvendada a estrutura do DNA (ácido desoxirribonucléico) por James Watson e Francis Crick[1, 2]; essa descoberta causou grande revolução científica, pois o DNA contém toda a informação genética de um indivíduo. O trabalho realizado por eles lhes rendeu o prêmio Nobel em medicina ou fisiologia.

Desde então, vários grupos de pesquisa têm buscado aprimorar o estudo do DNA e da sua decodificação. Em 1990, foi fundado o Projeto Genoma, com previsão de término para 2005 e cuja ideia básica era seqüenciar o genoma de organismos vivos. Como meta principal, era pretendido sequenciar todo o DNA do genoma humano, criando mapas e bancos de dados para a distribuição destas informações. Em 1998, foi formada uma organização internacional, que incluia os Estados Unidos, Austrália, Japão e alguns países da Europa, responsável por coordenar o Projeto HUGO (Human Genome Organisation)[3, 4]; projeto este voltado para a organização dos trabalhos com genoma humano em um banco de dados centralizado, o Genome Database, e a para a análise funcional do genoma, com aplicações voltadas para a medicina. O projeto HUGO, que foi parcialmente concluído em 2003, agora se volta para a disseminação de dados e de diretrizes para o estudo do genoma.

Existem outras organizações, análogas ao HUGO, para outras espécies. A SGD (*Saccharomyces Genome Database*) estuda a *Saccharomyces cerevisiae*[5], o Flybase trabalha com a *Drosophila melanogaster*[6], o TAIR com a *Arabidopsis thaliana*[7], o EcoCyc estuda a *Escherichia coli*[8], etc.

Sabemos que o sequenciamento do DNA não é suficiente para o entendimento do funcionamento celular. Precisamos entender os processos que transcrevem a informação contida no DNA em proteínas, responsáveis pelas reações bioquímicas nas células, e como estes são controlados e alterados. A partir daí podemos entender como essas reações formam o metabolismo celular.

Os processos que transformam a informação do DNA em proteínas são bem conhecidos. Primeiro ocorre a transcrição, quando é produzido RNA a partir de um pequeno trecho do DNA.

Toda a informação contida no trecho transcrito é carregada pelo RNA através do citoplasma até o ribossomo, onde ocorre a tradução, isto é, a produção de proteínas a partir de informação do RNA. As regiões do DNA que dão origem às proteínas são chamadas de genes[9].

À cadeia de DNA damos o nome de Genoma, sendo este constituído de genes, íntrons, éxons, etc. Ao conjunto de moléculas de RNA da transcrição damos o nome de Transcriptoma. O Proteoma é o conjunto das proteínas expressas pela célula. O nome dado ao conjunto de todos os componentes que podem participar das reações bioquímicas da célula (incluindo moléculas inorgânicas, nucleotídeos, etc) é Metaboloma. O Genoma de todas as células de um mesmo organismo deve ser idêntico, exceto por mutações somáticas, mas o Transcriptoma e o Proteoma, por serem expressões do DNA, podem variar em tecidos ou etapas diferentes dos processos celulares; assim, podemos obter informações momentâneas sobre as células.

Podemos descrever uma célula como uma rede formada com associações de diversos tipos, sendo estas entre proteínas, genes e metabólitos. A esta rede damos o nome de Interatoma[10]. Essas redes podem ser extremamente complexas, tornando difícil a previsão das respostas do sistema caso aconteçam modificações em algum dos elementos. Apesar de muito complexa, a análise do Interatoma nos permite fazer diagnósticos de performance metabólica, assim como estudar as alterações causadas pelo uso de fármacos e terapias, tornando-se extremamente útil para o estudo da genética.

A análise do Interatoma pode ser abordada de duas formas: podemos analisar cada componente separadamente, ou examinar a rede como um todo. Devido à existência de terapias bem sucedidas, envolvendo a manipulação de poucos genes, e ao estudo das características da rede[11], podemos inferir que existem módulos funcionais relacionados a funções biológicas específicas. Em compensação, a descrição do estado celular pela observação de um único gene não tem obtido resultados satisfatórios. Já foi mostrado que a análise de módulos autoassociados, mesmo considerando somente associações entre proteínas, quando estudados em conjunto com os dados de transcrição, pode descrever a performance celular e o estado metabólico das células[12–14].

1.1 Métodos de medida do metabolismo celular

Existem variados métodos de análise da performance celular. O Transcriptoma, conjunto de RNA de transcrição presente na célula, contém a informação de quais proteínas serão produzidas na célula, enquanto o Proteoma contém todas as proteínas desta. Entre as técnicas de quantificação do Transcriptoma e estão o microarranjo, o RNAseq e o PCR, este último

possuindo variações, que permitem a quantificação do Proteoma. Cada técnica tem as suas vantagens e desvantagens, sendo brevemente descritas a seguir.

A técnica de PCR[15] (*Polymerase Chain Reaction*) visa amplificar o sinal, ou quantidade, de um alvo, que pode ser DNA ou mRNA. O método consiste em fazer uma solução com a amostra, DNA polimerase e fragmentos de DNA complementares ao alvo desejado, chamados de *primers*, e realizar um ciclo térmico de aquecimento e resfriamento para quebrar as cadeias de DNA e depois reconstruí-las com os *primers*. Esta técnica amplifica exponencialmente o sinal dos alvos desejados, possibilitando a medida da quantidade de produtos presentes na solução com muita precisão. O grande problema desta técnica é que sua medida não é de genoma completo, analisando poucos RNAs; isso torna os dados obtidos por PCR pouco relevantes para os métodos desenvolvidos neste trabalho.

A técnica de microarranjo[16] (*microarray*) consiste em utilizar um *chip* com inúmeras sondas, com trechos de DNA, que se hibridizam, conectando-se a trechos de RNAs específicos. Estas sondas possuem fluorescência dependente da sua hibridização, portanto, quando um RNA se prende a uma sonda, esta emite luz. A partir da intensidade luminosa das sondas calcula-se a quantidade dos RNAs presentes na amostra. Esta técnica não é capaz de detectar RNAs que não sejam o alvo de nenhuma das sondas, limitando o experimento a RNAs conhecidos. Outro problema desta técnica é o grande ruído da medida, além de que uma mesma amostra examinada em laboratórios diferentes traz resultados muito diferentes, havendo necessidade de um estudo sobre a normalização destes dados. Apesar dos problemas da técnica, ela é muito difundida e grande parte dos dados disponíveis na rede foram obtidos a partir dela.

O RNAseq, ou seqüenciamento de RNA[17], é uma técnica de mapeamento e quantificação de Transcriptomas. É montada uma biblioteca com as seqüências de bases dos diferentes RNAs; é armazenada a sequüência de bases das cadeias de RNA formadas através de vários ciclos químicos. Então as seqüências de bases são identificadas na biblioteca e contadas, obtendo a quantidade de cada RNA da amostra com boa precisão. Esta técnica é recente, cara e pouco difundida, motivos pelos quais os dados utilizados neste trabalho não foram obtidos utilizando-a, mas é de se esperar que nos próximos anos o RNAseq se torne a técnica mais utilizada para a obtenção de dados de expressão gênica.

1.2 Transcriptograma

A medida do Transcriptoma através dos métodos citados na seção 1.1 nos mostra quantitativamente a atividade dos processos bioquímicos. Com essa análise, podemos observar as

variações no Transcriptoma causadas por patologias e terapias; assim como encontrar novos alvos farmacológicos e direcionar as pesquisas de desenvolvimento de fármacos. Podemos, também, usar esta análise para a observação de marcadores de doenças e diagnósticos[18–21].

O problema do estudo de expressão gênica a partir dos experimentos de microarranjo surge do fato de que o ruído da medida é muito grande. Para reduzir esse erro, devemos desenvolver um método que aumente a razão entre sinal e ruído, aumentando a qualidade da análise dos dados de microarranjo.

Em seus trabalhos, Rybarczyk et al.[12–14] propuseram um método computacional de ordenamento do proteoma, e análise de Transcriptoma, capaz de criar uma lista de proteínas em que a proximidade de duas proteínas é relacionada à probabilidade de haver associação entre elas. Havendo esta relação entre proteínas próximas, as regiões passam a ter significado e pode-se tomar médias locais, que reduzem o efeito de ruído.

A análise de Transcriptoma tomando médias locais sobre a rede ordenada, como sugerido por Rybarczyk[12, 13], é chamada Transcriptograma. Sob esta ótica, deixa-se de trabalhar com uma proteína e passa-se a trabalhar com um grupo destas. Este método e suas aplicações, como o diagnóstico proposto por Benetti[14], são a base deste trabalho e serão explicados com detalhes no capítulo 3. O estudo do Proteoma ainda não é possível da forma que fazemos para o Transcriptoma porque não foram desenvolvidas técnicas de medida de expressão gênica de genoma completo para proteínas, mas havendo novas técnicas, é possível utilizar os mesmos métodos de análise de dados para o Proteoma.

1.3 Objetivos

O objetivo deste trabalho é propor uma generalização da análise da expressão gênica realizada por Rybarczyk e Benetti[12–14]. Enquanto naqueles trabalhos o Proteoma foi ordenado em uma dimensão, levando a bons resultados e demonstrando o potencial da técnica, neste trabalho pretendo mostrar uma ferramenta capaz de organizar e analisar a rede em duas dimensões.

A mudança de uma para duas dimensões reduz os problemas de frustração da rede, possibilitando o agrupamento mais coerente dos módulos da rede e aumentando a relação entre a proximidade das proteínas e a probabilidade de haver associação entre elas. A melhora no ordenamento resulta em um transcriptograma mais correto, pois quanto mais direta a relação entre a proximidade dos nós da rede e a probabilidade de que estes nós representem proteínas associadas, mais sentido existe em tomar médias que relacionam proteínas próximas.

A partir do transcriptograma pretendo propor um método de diagnóstico puramente matemático. Para isso, criamos os padrões para as duas classes que queremos diagnosticar a partir de um conjunto de amostras com classificação conhecida, encontrando os perfis de média e desvio padrão do transcriptograma de cada classe. Possuindo os padrões, analisamos o transcriptograma de uma amostra comparando com as classes e calculamos a probabilidade dele pertencer a estas.

2 *Revisão*

Neste capítulo, abordamos a representação de sistemas na forma de redes, alguns modelos de redes existentes e descrevemos alguns métodos utilizados para ordenar estas. Descrevemos também as fontes dos dados e as ferramentas, disponíveis na *internet*, utilizados neste trabalho.

2.1 Redes

Uma rede é um modelo de representação constituído por vértices, ou nós, e arestas. Cada nó representa um item da rede, enquanto as arestas representam interações ou conexões entre estes itens. Esta modelagem pode representar uma gama imensa de sistemas, como redes sociais, biológicas, de informação, tecnológicas, etc.

As redes sociais podem ser representações de grupos de pessoas onde cada indivíduo é um nó, e as suas interações (relações familiares, afetivas e profissionais, entre outras) são as arestas, que ligam os nós. As redes tecnológicas descrevem sistemas criados pelo homem, como redes de estradas ou linhas aéreas. Redes biológicas são as que descrevem relações biológicas, como uma teia alimentar, redes neurais, sistema vascular, etc. Em resumo, uma rede é uma representação simplificada de um sistema de indivíduos relacionados; existem modelos matemáticos para a análise destas redes[22–24].

O Interatoma pode ser representado por uma rede, onde cada proteína é um nó e as suas associações são as arestas. Neste trabalho consideramos que as associações não possuem direção. Podemos, então, criar uma matriz de Adjacência A, onde a cada elemento a_{ij} é atribuído um valor 0 ou 1, representando a ausência ou existência, respectivamente, de interação entre os nós i e j. Este modo de descrição é bem representado na figura 2.1.

2.1.1 Medidas das propriedades da rede

A matemática de redes propõe medidas das qualidades e características desses sistemas. Algumas destas medidas são apresentadas a seguir.

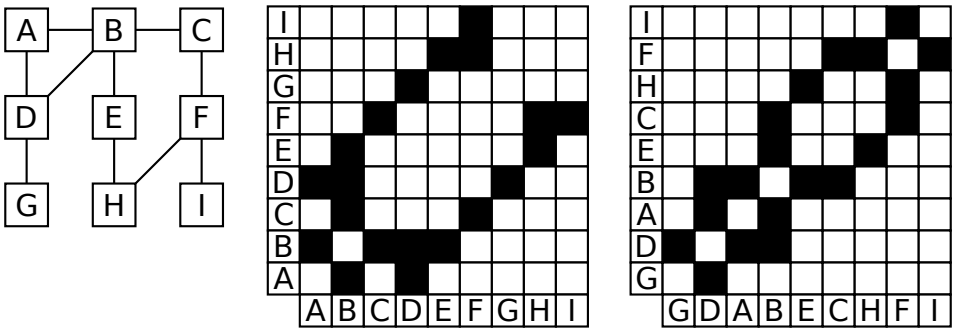

Figura 2.1: Rede simples representada por seus nós e arestas e duas possíveis matrizes de adjacência, onde o espaço preto representa interação e o branco, ausência desta; ou seja, 1 e 0, respectivamente.

Distribuição de conectividades

A conectividade k_i representa o número de ligações das quais o vértice i faz parte. Pode ser calculada a partir da matriz de adjacência:

$$k_i = \sum_{j}^{N} a_{ij} \tag{2.1}$$

onde N é o número total de vértices e a_{ij} é o elemento da matriz de Adjacência. Para definir a conectividade média da rede, basta somar todos os elementos k_i e dividir por N:

$$\langle k \rangle = \frac{1}{N} \sum_{i}^{N} k_i \tag{2.2}$$

A distribuição de conectividades $p(k)$ representa a fração de genes com conectividade k. Esta pode ser calculada a partir da conectividade:

$$p(k) = \frac{1}{N} \sum_{i}^{N} \delta_{kk_i} \tag{2.3}$$

Coeficiente de clusterização

O coeficiente de clusterização C_i é uma medida da aglomeração dos nós da rede; representa a fração das ligações entre os vértices ligados ao nó i em relação ao total de ligações possíveis entre esses nós. Podemos calcular esse coeficiente a partir da conectividade:

$$C_i = \frac{2n}{k_i(k_i-1)} = \frac{1}{k_i(k_i-1)} \sum_{j}^{N} (a_{ij} \sum_{m}^{N} a_{jm} a_{mi}) \tag{2.4}$$

onde n é o número de ligações entre os vizinhos do nó i. C_i pode variar de 0 a 1, representando que os nós ligados a i não têm nenhuma ligação ou que são totalmente ligados entre si, respectivamente. Podemos calcular a média da clusterização assim como calculamos a da conectividade:

$$\langle C \rangle = \frac{1}{N} \sum_{i}^{N} C_i \tag{2.5}$$

Overlap topológico

O *overlap* topológico é uma medida proposta por Barabási e Ravasz[25, 26] que representa o grau de compartilhamento de vizinhança entre dois vértices. Enquanto a medida de clusterização é representada por um tensor de primeira ordem, ou vetor, pois se refere aos vértices, representa-se o *overlap* topológico com um tensor de segunda ordem, ou matriz, pois este compara dois nós. Essa matriz pode ter valores diferentes de zero para vértices não ligados, desde que os seus vizinhos tenham ligações entre si. Os elementos da matriz de *overlap* topológico O_{ij} podem ser calculados da seguinte forma:

$$O_{ij} = \frac{a_{ij} + \sum_{m}^{N} a_{im} a_{jm}}{min(k_i, k_j) + 1 - a_{ij}} \tag{2.6}$$

onde o numerador representa o número de nós aos quais i e j estão ligados mais a_{ij}, que é o elemento da matriz de adjacência; no denominador, $min(k_i, k_j)$ é o menor entre k_i e k_j, e k_i é a conectividade do nó i.

2.1.2 Tipos de redes

Rede aleatória

O modelo de rede aleatória, criado em 1960 por Erdös-Rényi[27], sugere uma rede de N vértices com probabilidade p de haver uma ligação entre cada par de vértices. Sob essas condições, é gerada uma rede com aproximadamente $pN(N-1)/2$ arestas, cuja distribuição de conectividades segue uma lei de distribuição binomial, havendo uma grande quantidade de nós com conectividade próxima da média $p(N-1)/2$ e poucos nós com conectividades altas ou baixas. A clusterização deste tipo de rede se mostra independente da conectividade, sendo praticamente constante e da ordem de p.

Rede livre de escala

A rede livre de escala é formada por um modelo de crescimento de redes proposto por Barabási e colaboradores[29]. O modelo de crescimento propõe que a rede seja iniciada com poucos nós, todos interligados, e que sejam acrescentados novos vértices um a um. A probabilidade de haver ligações entre o novo vértice e um vértice antigo é dada pela seguinte equação:

$$p = \frac{k_i}{\sum_j^N k_j} k_0 \tag{2.7}$$

onde k_i é a conectividade do vértice i e k_0 é a conectividade média desejada para a rede.

O processo descrito acima gera uma rede com distribuição de conectividade $p(k) \sim k^{-\gamma}$ na forma de lei de potência caracterizada por $2 < \gamma < 3$. Devido à distribuição das probabilidades de novas ligações serem maiores para nós mais conectados, a probabilidade de haver vértices altamente conectados é alta, estes são chamados *hubs*. Outra característica desta rede é a distribuição de clusterização homogênea, independente da conectividade, indicando que não há formação de módulos.

Rede hierárquica

Podemos construir uma rede hierárquica[30] a partir de um bloco com N vértices, todos ligados entre si, replicando esse bloco m vezes; e então ligando o nó central de cada novo bloco ao nó central do bloco inicial. Esse processo pode ser realizado várias vezes para alcançar o tamanho de rede desejado.

Esse modelo de rede é livre de escala, mas apresenta estrutura modular. a distribuição de conectividades mostra-se como uma lei de potências com $\beta = 2,26$; assim como o coeficiente de clusterização, que segue $C(k) \sim k^{-\beta}$. Podemos ver que existem poucos nós muito conectados e muitos nós de baixa conectividade; os primeiros são chamados "hubs", com baixa clusterização, enquanto os outros (menos conectados) fazem parte de pequenos módulos com alta clusterização.

2.2 Métodos de ordenamento

O ordenamento é uma forma de classificação de redes que visa a organização destas. Os métodos de ordenamento se aplicam quando há uma enorme quantidade de dados e queremos juntar os nós em variados grupos, possibilitando a análise da rede através destes. O ordenamento

é o primeiro passo para a análise do Proteoma e do Transcriptoma, assim como para a criação do Transcriptograma; portanto, é essencial estudar e aprimorar os métodos de ordenamento de redes.

2.2.1 Agrupamento hierárquico

O método de agrupamento hierárquico proposto por Barabási e colaboradores[25, 26] foi inspirado no método de formação de dendogramas[28], diagramas em forma de árvore que agrupam itens similares a partir de uma matriz de proximidade. Barabási usou a matriz de *overlap* topológico como critério de semlhança para montar um diagrama de agrupamento em que genes de um mesmo grupo possuem vizinhos interligados.

O algoritmo para o agrupamento hierárquico de N genes ou proteínas parte da matriz de *overlap* topológico destes nós. Encontramos o par de nós com maior valor de *overlap*, digamos (a,b), e a seguir montamos uma nova rede, com $N-1$ vértices, onde o par selecionado é eliminado e substituído por um único nó, indexado como (a,b). Os elementos da nova matriz podem ser obtidos a partir do seguinte cálculo:

$$O_{(a,b)c} = \frac{k_a O_{ac} + k_b O_{bc}}{k_a + k_b} \tag{2.8}$$

onde k_a é a conectividade do vértice a e O_{ac} é o valor de *overlap* topológico entre os nós a e c. O processo se repete até que se forme uma rede com apenas um nó. Um resultado possível de uma rede com 8 nós seria $((((A,(D,F)),(B,C)),(E,H)),G)$. Nesta representação, pares dentro de parênteses possuem mais similaridades entre si do que com os nós de outros índices; portanto, nós vizinhos devem possuir alto *overlap* topológico.

A vizinhança de um gene, neste diagrama, tem significado, visto que os vértices foram ordenados de acordo com o seu *overlap* topológico. Também podemos representar este ordenamento como uma lista de proteínas; perdendo a informação de hierarquia, mas mantendo o ordenamento, teríamos, para a rede anterior, o seguinte resultado: $ADFBCEHG$. Seguindo este método, podemos obter 2^{N-1} ordenamentos diferentes e equivalentes, visto que para cada vez que realizamos o processo de agrupamento, os nós A e B podem se juntar na forma (A,B) ou (B,A).

2.2.2 Minimização da função de custo

O método da minimização da função de custo, proposto por Rybarczyk e colaboradores[12, 13], se aplica a redes binárias não direcionadas. O objetivo final deste ordenamento é classificar a

lista de genes/proteínas de forma que os nós associados se aproximem, tornando possível a observação de módulos fortemente ligados; também se espera que módulos que conectam-se fiquem próximos.

A implementação do método parte da matriz de adjacência descrita anteriormente, que é representada como na figura 2.1, onde cada ponto preto significa a existência de associação e o branco, ausência desta. É proposta uma função custo que penalize dois fatores: a distância da ligação para a diagonal da matriz de adjacência e a quantidade de interfaces ao redor de uma ligação, que representa, para dois nós ligados, quantos dos vizinhos do primeiro não interagem com o segundo e vice-versa. A função custo total é descrita pela equação 2.9. As posições dos nós são trocadas aleatoriamente, sendo aceitas aquelas que diminuem a função custo. Para evitar que o ordenamento fique estagnado em uma situação metaestável de alta energia, é utilizado um método de *simulated annealing* que inclui um parâmetro semelhante à temperatura, que é reduzido ao longo do processo. O resultado do método é uma configuração estável que minimiza o custo total.

$$\varepsilon = \sum_{i,j=1}^{N} \varepsilon_{i,j} = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}|i-j|^{\alpha}(4 - a_{i,j+1} - a_{i,j-1} - a_{i+1,j} - a_{i-1,j}) \tag{2.9}$$

É importante observar que, no processo, associações não são criadas nem destruídas; o método apenas organiza a matriz de adjacência. Na figura 2.2 observa-se a evolução temporal, em passos de Monte Carlo (um passo de Monte Carlo equivale a N testes de troca), da matriz de adjacência de uma rede simples com dois módulos muito "clusterizados". Mesmo nos ordenamentos finais, existem pontos distantes da diagonal e espaços brancos dentro dos módulos; estes ocorrem porque os nós, mesmo pertencendo a um módulo, possuem algumas interações com vértices do outro e não se conectam com todos os vértices de seu módulo.

Na figura 2.3 podem ser observados os estados finais das matrizes de adjacência das redes de proteínas de alguns organismos. Nestes ordenamentos também observamos os mesmos defeitos da rede de dois módulos, mas em maior quantidade; isso ocorre devido à complexidade das redes naturais, que possuem módulos menos "clusterizados" e com mais ligações externas. Estes defeitos são chamados frustrações da rede.

Este método de ordenamento, como demonstrado por Rybarczyk, Benetti et al.[12–14], quando aplicado ao Interatoma, separa as proteínas em módulos e estes caracterizam funções biológicas, como visto na figura 2.4. Este resultado permitiu fazer observações mais precisas de performance celular e aumentou o poder de diagnóstico da análise de microarranjo; pois as médias locais do transcriptoma no sistema ordenado aumentaram a razão sinal-ruído, como veremos

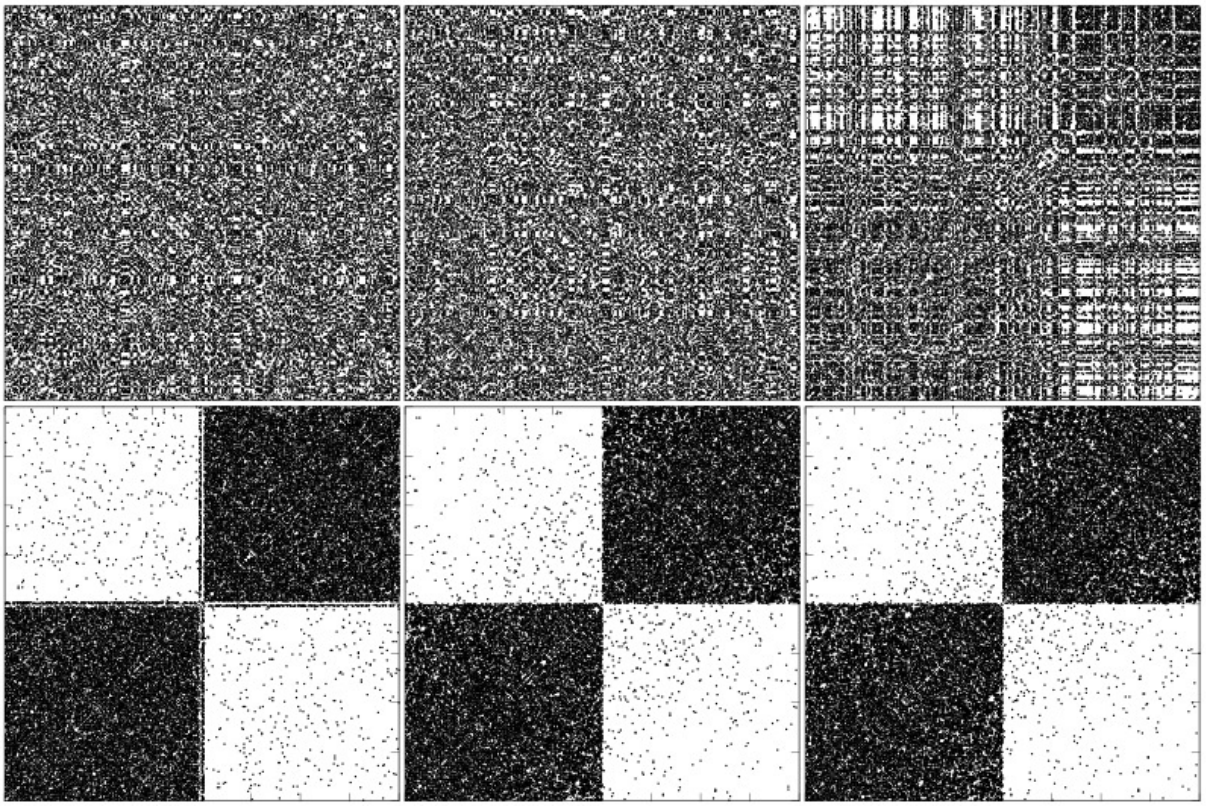

Figura 2.2: Evolução do ordenamento de uma rede simples em uma dimensão com dois módulos após 0, 600, 1200, 1800, 2400 e 3000 passos de Monte Carlo.

mais adiante.

2.3 Bancos de dados

Existem muitos dados de associação proteica, obtidos por diferentes laboratórios e grupos de pesquisa, estes dados são disponibilizados em vários bancos de dados, que são iniciativas nacionais e supra nacionais. Estes grupos possuem colaboração e a quantidade de informações é enorme. O problemas reside na dificuldade da interpretação desta quantidade de dados, sendo complexo obter informações a partir de tantos resultados sem simplificar demais este sistema.

2.3.1 STRING

O STRING (*Search Tool for the Retrieval of Interacting Genes/proteins*) é um banco de dados contendo as associações conhecidas e previstas entre genes ou proteínas. Na versão atual, *STRING*9.0, estão incluidos mais de 5 milhões de proteinas, para mais de mil organismos[31–34]. Entre os mantenedores deste banco de dados, estão o CPR (*NNR Center for Protein Research*), o EMBL (*European Molecular Biology Laboratory*), o SIB (*Swiss Institute of Bioinformatics*), o

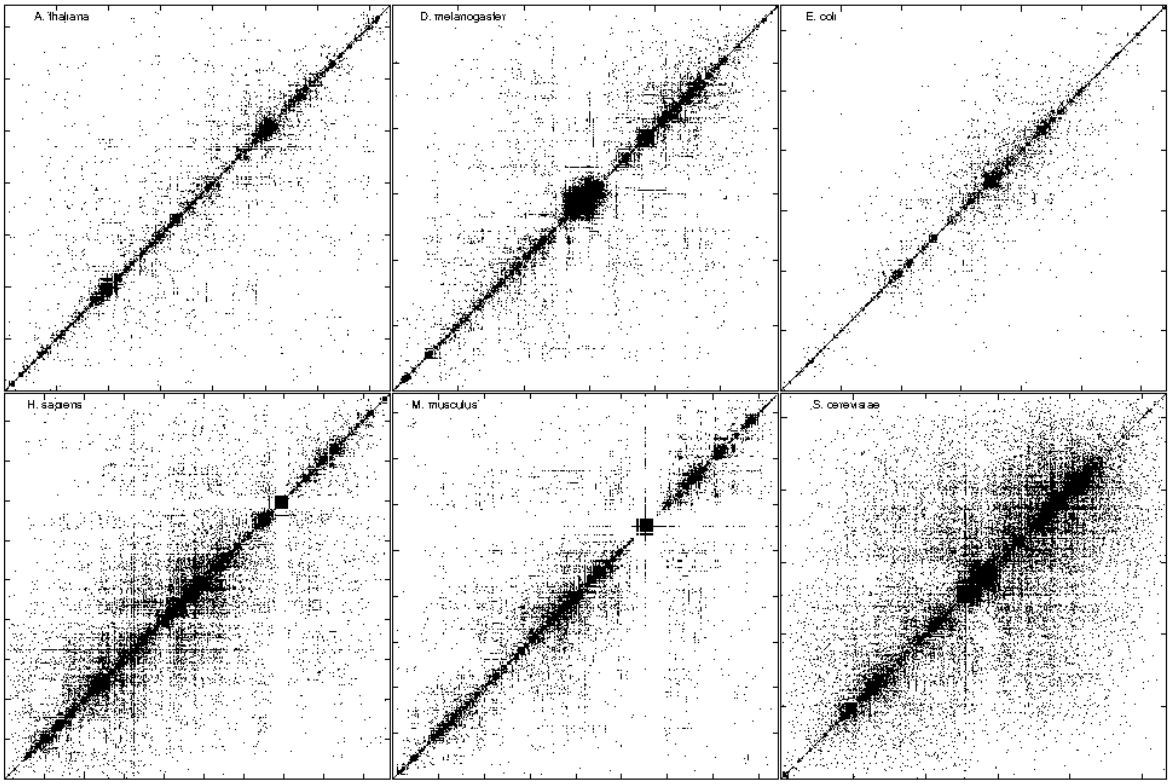

Figura 2.3: Matrizes de adjacência para os ordenamentos finais em uma dimensão para os organismos *Arabidopsis thaliana*, *Drosophila melanogaster*, *Escherichia coli*, *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae*.

SUND-KU (*University of Copenhagen*), o TUD (*Technical University Dresden, Biotec*) e UZH (*University of Zurich*). Além das pesquisas destes grupos, o STRING utiliza outros bancos de dados como fontes, descritas na tabela 2.1.

O STRING disponibiliza dados de associação proteína-proteína com um grau de confiabilidade que pode ser ajustado manualmente, possibilitando o controle de falsos positivos e negativos com maior precisão. A partir dos dados de associação proteína-proteína de um organismo, podemos criar a rede de associação, possibilitando a abordagem matemática utilizada neste trabalho.

Os dados de associação proteica do STRING são classificados segundo duas categorias: associações físicas (diretas) e associativas, de rotas metabólicas. Dentro destas categorias temos mais quatro subdivisões, segundo os tipos de informação: alta-performance, coexpressão, conhecimento prévio e contexto genômico. Os sete métodos utilizados para a obtenção destes dados são: *co-expression*, *co-occurrence*, *database*, *experiments*, *gene fusion*, *neighborhood* e *textmining*. As associações proteicas propostas são avaliadas analisando a probabilidade das duas proteínas em questão participarem de uma mesma rota metabólica do KEGG (*Kyoto Encyclopedia of Genes and Genomes*)[35], sítio que contém dados sobre rotas metabólicas com alto

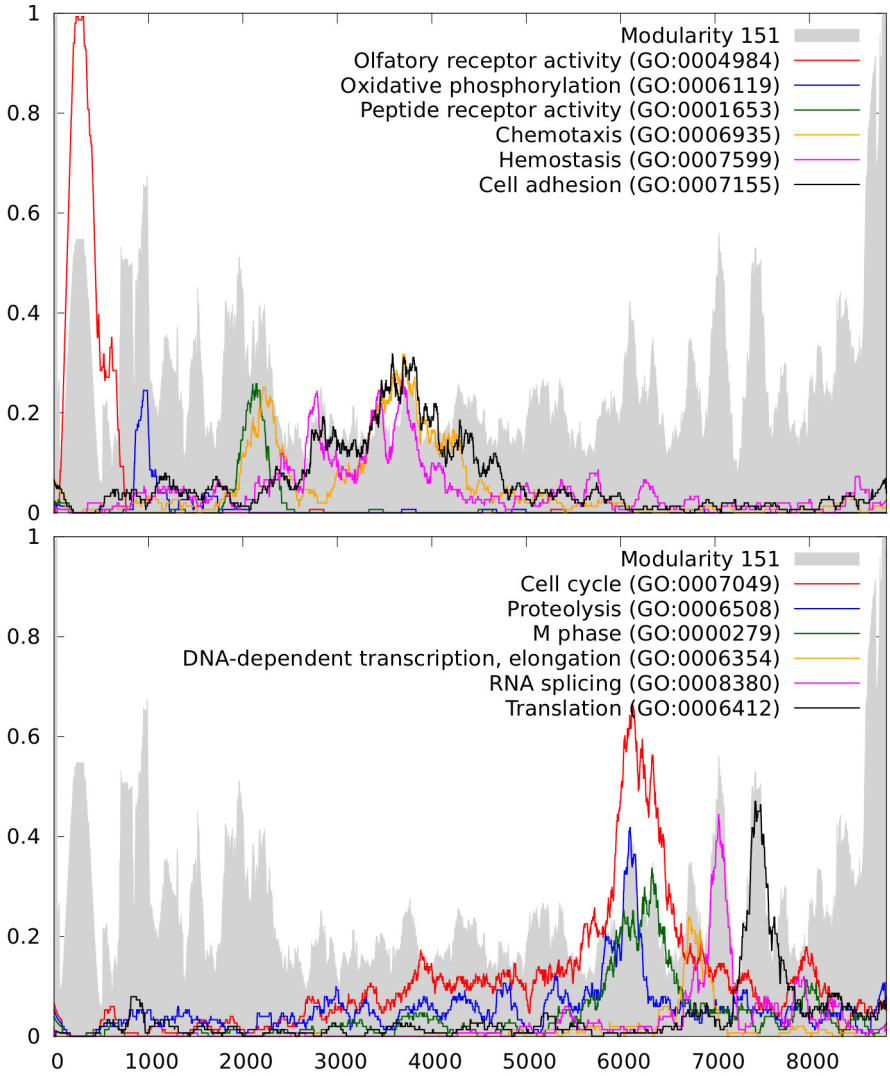

Figura 2.4: Distribuição de alguns termos de ontologias no ordenamento de *Homo sapiens* em uma dimensão.

grau de confiança, pois todos os seus dados são curados manualmente. É possível escolher quais destes métodos são considerados ao construir a rede e, dependendo do objetivo do usuário, os critérios escolhidos para a associação de proteínas podem variar. As classificações são explicadas a seguir.

- ***Neighborhood*, *Gene Fusion* e *Co-Occurrence***: pares que funcionam de maneira associada devido à pressão seletiva envolvida no processo de evolução.
- ***Co-Expression***: pares cuja probabilidade de co-expressão em um mesmo organismo é elevada.
- ***Experiments***: associaçcões retiradas de bancos de dados experimentais.
- ***Databases***: pares deduzidos por especialistas a partir de bancos de dados.

Tabela 2.1: Bancos de dados que fazem parte do STRING

Sigla	Nome do banco de dados	Sítio
COG	*Clusters of Orthologous Groups*	www.ncbi.nlm.nih.gov/COG/
Ensembl	*Ensembl*	www.ensembl.org/
IntAct	*Intact*	www.ebi.ac.uk/intact/
RefSeq	*Reference Sequence*	www.ncbi.nlm.nih.gov/RefSeq/
PubMed	*US National library of Medicine*	www.ncbi.nlm.nih.gov/pubmed/
REACTOME	*Reactome*	www.reactome.org/
DIP	*Database of Interacting Proteins*	dip.doe-mbi.ucla.edu/
BioGRID	*Biological General Repository for Interaction Datasets*	thebiogrid.org/
MINT	*Molecular INTeraction Database*	mint.bio.uniroma2.it/
KEGG	*Kyoto Encyclopedia of Genes and Genomes*	www.genome.jp/kegg/
SGD	*Saccharomyces Genome Database*	www.yeastgenome.org/
FlyBase	*Database of Drosophila Genes & Genomes*	flybase.org/
UniProt	*Uniprot*	www.uniprot.org/
SwissModel	*SWISS-MODEL*	swissmodel.expasy.org/
HUGO	*Human Genome Organisation*	www.hugo-international.org/
OMIM	*Online Mendelian Inheritance in Man*	www.omim.org
NCI/ Nature PID	*National Cancer Institute / Nature Pathway Interaction Database*	pid.nci.nih.gov/
PDB	*Protein Data Bank*	www.rcsb.org/pdb/
TIF	*The Interactive Fly*	www.sdbonline.org/fly/ aimain/1aahome.htm
BioCyc	*BioCyc Database Collection*	www.biocyc.org/
GO	*Gene Ontology*	www.geneontology.org/
SIMAP	*The Similarity Matrix of Proteins*	boincsimap.org

- ***Textmining***: associa proteínas citadas no resumo de um mesmo artigo do PubMed.

Para cada par de proteínas e para cada um destes métodos, o STRING atribui um valor de confiabilidade S_i; aplicando a equação 2.10 a esses valores, é montado o *score* final. As redes utilizadas neste trabalho foram criadas com as associações com *score* superior a $0,8$, desconsiderando *textmining*. O método de *textmining* foi desconsiderado porque não trata de associação real entre as proteínas, mas do fato delas aparecerem em um mesmo artigo.

$$S = 1 - \prod_{i=1}^{7}(1 - S_i) \tag{2.10}$$

2.3.2 GO - *The Gene Ontology*

O banco de dados *Gene Ontology*[36], disponível no sítio *http://www.geneontology.org/*, tem como objetivo organizar os dados sobre os genes e seus produtos, assim como padronizar estas informações. O GO possui informações sobre mais de 50 espécies, organizadas em três classificações diferentes, chamadas ontologias. Cada gene e seus produtos é relacionado às diferentes ontologias, que são independentes dos organismos. A organização das ontologias se dá hierarquicamente, cada uma com suas ancestrais, que as incluem, e filhas, subdivisões dessa classe. As três ontologias são diferentes porque usam critérios distintos de classificação, e são descritas a seguir.

- ***Cellular Component***: Os genes são classificados de acordo com a componente celular onde os seus produtos agem; esta classificação não se refere às funções exercidas pelos produtos dos genes nem aos processos biológicos dos quais eles participam. A ontologia GO:0016020, por exemplo, contém 14192 genes e seus produtos, localizados na membrana celular.

- ***Molecular Function***: Os genes são classificados segundo as atividades exercidas por seus produtos em escala molecular; desconsiderando os contextos nos quais estas funções estão inseridas. A ontologia GO:0016209, por exemplo, contém 130 genes e produtos responsáveis pela atividade antioxidante.

- ***Biological Function***: Os genes são classificados de acordo com os processos biológicos dos quais seus produtos fazem parte, independentemente das funções moleculares exercidas pelos seus produtos e das componentes celulares onde eles agem. A ontologia GO:0044237 contém 15223 genes e produtos referentes ao processo metabólico celular.

No mesmo sítio, está disponível a ferramenta AmiGO[37], com a sua função *Term Enrichment*, que nos possibilita calcular o que é conhecido como enriquecimento funcional das proteínas. Podemos entrar com uma lista de proteínas e um *background*, e ele nos responde as ontologias mais expressas naquele grupo; o que facilita o estudo da distribuição das funções na rede.

2.3.3 *GEO - Gene Expression Omnibus*

O *Gene Expression Omnibus*[38] é um banco de dados que comporta e distribui gratuitamente dados de transcriptomas, realizados com microarranjo e outros experimentos de alto

rendimento, submetidos pela comunidade científica. Os dados experimentais fornecidos pelo GEO são extremamente importantes para este trabalho, pois a partir do transcriptoma fazemos o transcriptograma, que é uma das finalidades deste trabalho e base da proposta de diagnósticos que será demonstrada.

3 *Organização de redes*

Neste capítulo é discutido o desenvolvimento do método de ordenamento de redes via minimização da função custo, em uma e duas dimensões. Primeiro será apresentadp com detalhes o processo desenvolvido nos trabalhos de Rybarczyk et al.[12–14], para então apresentar as propostas deste trabalho referentes ao ordenamento de redes, que incluem a melhora da técnica de MFC existente e a generalização do método para mais de uma dimensão.

3.1 Método de minimização da função custo (MFC)

O método de minimização da função custo, como descrito superficialmente no capítulo 2, consiste em ordenar a matriz de adjacência, que é uma matriz muito específica. Sua característica principal é que $a_{ij} = a_{ji}$, onde a_{ij} pode assumir apenas os valores 1 ou 0. Sendo assim, esta matriz pode representar qualquer rede binária com ligações não direcionadas. Foi considerado que um vértice não interage com ele mesmo, portanto $a_{ii} = 0$.

Dada uma rede R com N nós e L ligações, indexamos aleatoriamente os vértices com números entre 1 e N. Montamos a matriz de adjacência A com N linhas e N colunas, com a_{ij} igual a 1 se houver interação entre i e j e 0 se esta interação não existe. Esta matriz é facilmente representada como na figura 2.1, onde os sítios pretos representam $a_{ij} = 1$, e os brancos representam $a_{ij} = 0$.

Podemos ter $N!$ matrizes de adjacência diferentes, devido à possibilidade de rotulações diferentes, e queremos encontrar qual destas possui o ordenamento unidimensional que melhor agrupa os módulos da rede. Podemos procurar uma matriz que possua uma maior densidade de elementos não nulos próximos à diagonal, maximizando o número de pares interagentes posicionados como vizinhos diretos; também podemos buscar uma matriz que reduza interfaces na representação gráfica da matriz de adjacência, dando preferência para matrizes em que o elemento se conecte com os vizinhos dos elementos ligados a ele, formando agrupamentos de elementos ligados; mas estas duas possibilidades de ordenamento nem sempre são concordan-

tes. Estas condições são levadas em conta na função custo proposta pela equação 2.9 e vão competir quando tentamos minimizar este custo.

O método utilizado para realizar a escolha da matriz de adjacência ótima é um método de Monte Carlo. Partindo de uma matriz de adjacência, calculamos o custo total ε_0, dado pela equação 2.9, e sorteamos dois nós. Trocamos a posição destes nós, ou seja, trocamos suas linhas e colunas; e calculamos o novo custo ε_f. Se a troca reduz o custo, $\Delta\varepsilon < 0$, é aceita a nova matriz de adjacência; se a troca aumentar o custo, $\Delta\varepsilon > 0$, aceitamos esta com uma probabilidade $P = e^{(-\frac{\Delta\varepsilon}{T})}$, onde T é um parâmetro semelhante à temperatura do sistema. Este aceite de uma troca negativa é uma técnica utilizada para evitar que o processo fique preso em uma configuração de mínimo local da função custo. Utilizamos a técnica de *simulated annealing*, baseada no resfriamento lento de metais com o objetivo de torná-los mais homogêneos e com menos frustrações e tensões; nesta técnica, o parâmetro T é inicialmente alto e é reduzido lentamente em forma de degraus ao longo do processo, de forma que o sistema atinja o equiliíbrio termodinâmico. O tempo, ou número de passos, adequado para parar a simulação seria aquele em que, com a temperatura suficientemente baixa, não se aceitam mais trocas. Ao final do processo, o sistema deve estar ordenado da forma que melhor concorde com as condições propostas e com o mínimo de frustrações da rede.

3.1.1 Redes artificiais

Aqui são propostas duas redes para testar o método, ambas pequenas e simples. Estas redes foram criadas para analisar pontos específicos do processo de ordenamento em uma dimensão.

A primeira rede possui dois módulos distintos de 200 nós cada, totalizando $N = 400$. A probabilidade de interação dos vértices é descrita na tabela 3.1. Esta rede gera a matriz de adjacência mostrada na figura 3.1, onde o primeiro módulo está posicionado com os índices de 1 a 200 e o segundo de 201 a 400.

Tabela 3.1: Probabilidades percentuais de interação entre nós dos módulos da primeira rede.

	Módulo I	Módulo II
Módulo I	50%	1%
Módulo II	1%	50%

A segunda rede possui quatro módulos. Destes, os módulos I e IV possuem 130 vértices cada e os outros têm 70 cada, totalizando $N = 400$. A probabilidade de interação entre dois vértices são descritas na tabela 3.2. A matriz de adjacência desta rede pode ser observada na figura 3.2. Esta rede vai ser útil na observação de frustrações do ordenamento em uma dimensão.

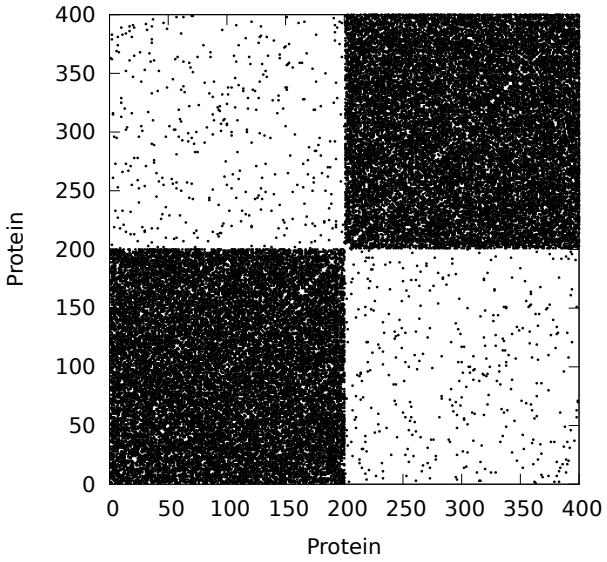

Figura 3.1: Matriz de adjacência de uma rede simples com dois módulos.

Tabela 3.2: Probabilidades percentuais de interação entre nós dos módulos da segunda rede.

	Módulo I	Módulo II	Módulo III	Módulo IV
Módulo I	60%	20%	20%	10%
Módulo II	20%	40%	10%	20%
Módulo III	20%	10%	40%	20%
Módulo IV	10%	20%	20%	60%

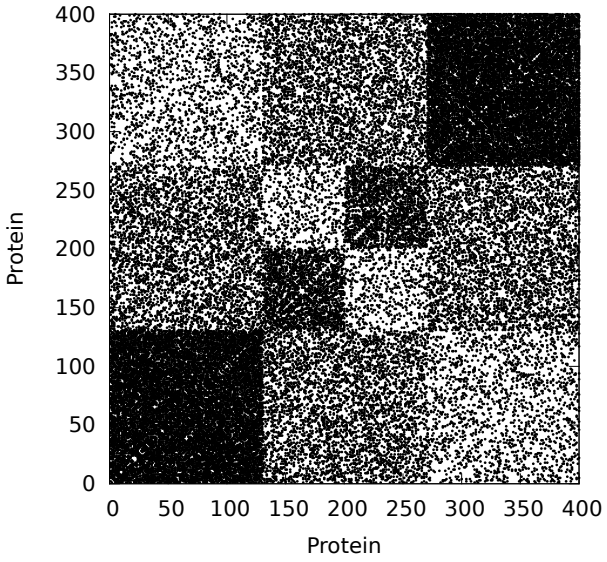

Figura 3.2: Matriz de adjacência de uma rede simples com quatro módulos.

O resultado do ordenamento em uma dimensão utilizando o MFC deveria separar os módulos; como obtido para a primeira rede, cuja evolução temporal está descrita na figura 3.3. O resultado obtido com a segunda rede e descrito na figura 3.4, no entanto, não é tão bom quanto o anterior.

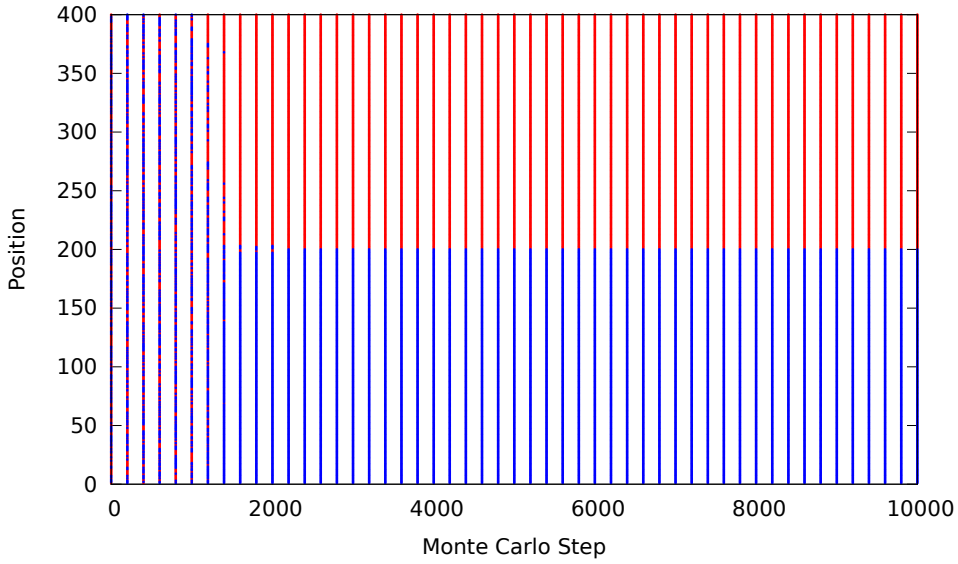

Figura 3.3: Evolução do ordenamento da primeira rede, onde os pontos vermelhos pertencem ao módulo I e os azuis ao módulo II.

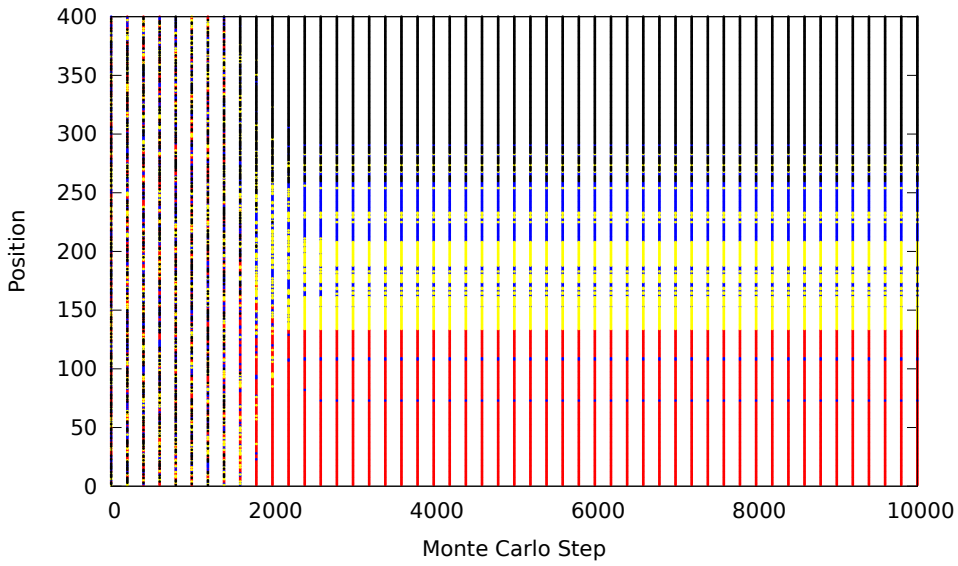

Figura 3.4: Evolução do ordenamento da segunda rede, onde os pontos vermelhos pertencem ao módulo I, os azuis ao II, os amarelos ao III e os pretos ao IV.

O problema que ficou evidente no ordenamento da segunda rede é a frustração geométrica, que ocorre ao tentarmos representar uma rede complexa em uma dimensão. É impossível encontrar um ordenamento que minimize o custo para todos os nós da rede, pois o ordenamento que resolve o custo para um vértice penaliza outro. Dentro dos módulos ocorrem frustrações. Até mesmo em sistemas muito simples, qualquer rede em que um nó possua mais de duas ligações, surgirá alguma frustração, mesmo que muito pequena.

3.1.2 Redes de proteínas

Aqui são ordenados os Proteomas de alguns dos organismos mais estudados: *Arabidopsis thaliana*, *Drosophila melanogaster*, *Escherichia coli*, *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae*. Estes organismos são escolhidos como bons modelos pois são as redes mais completas e com mais confiança que obtemos no STRING.

Obtemos as redes de proteínas a partir do STRING. Dentre os sete métodos para inferir associação entre proteínas fornecidos pelo banco, utilizamos seis, excluindo apenas o *text mining*. Selecionamos, entre todas as associações proteicas indicadas pelo STRING para um mesmo organismo, apenas aquelas que obtiveram escore S acima de $0,8$, calculado a partir da equação 2.10. Para ignorar o método de *textmining*, basta considerar o seu score parcial, S_7, igual a zero para todas as associações.

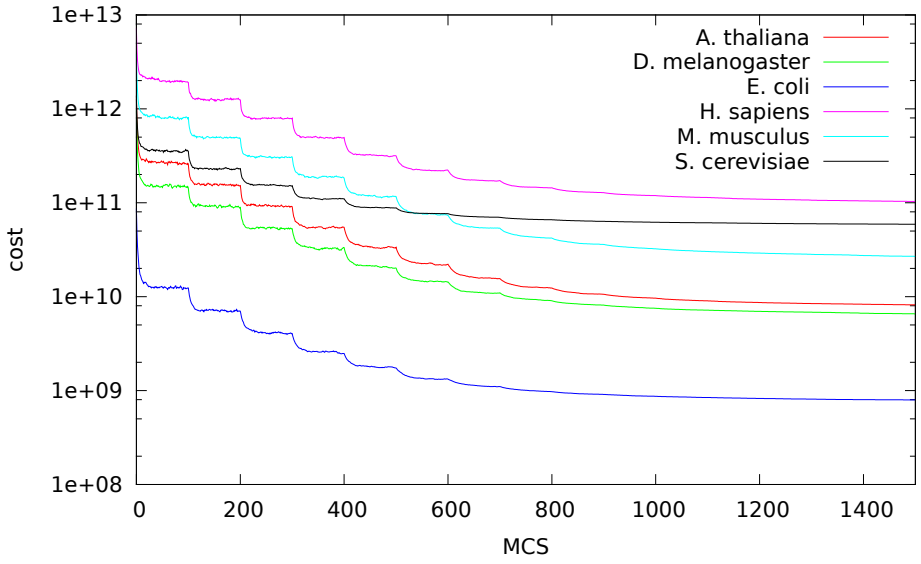

Figura 3.5: Custo total ao longo do ordenamento para variados organismos. A temperatura inicial é de $T_0 = 0,0001\varepsilon$. A cada cem passos de Monte Carlo a temperatura foi reduzida pela metade. Foram realizados dez mil passos de Monte Carlo.

Os organismos foram ordenados e seus módulos foram agrupados. A evolução da função custo total ao longo do processo é descrita pela figura 3.5, na qual não são mostrados os passos de 1500 a 10000 porque a variação do custo total é pequena, não agregando valor ao gráfico, mas ainda havendo muitas trocas relevantes para o resultado final. Os estados finais podem ser observados na figura 2.3, na qual fica claro que existem muitas frustrações na rede. Apesar das frustrações, o ordenamento parece efetivo ao separar as ontologias de funções biológicas, como pode ser observado na figura 2.4. O maior problema deste método está no seu custo computacional que, apesar de muito menor do que o de métodos de Dinâmica Molecular, é grande para redes com muitos nós e ligações, podendo levar mais de um mês para se ordenar

uma rede natural.

3.2 Otimização computacional do método MFC

O método de MFC é muito interessante e possibilita o ordenamento de redes grandes, com milhares de nós, mas ainda é computacionalmente custoso. Alguns pontos simples podem acelerar o processo.

A primeira operação que pode ser otimizada é o cálculo da função custo, que é feita duas vezes a cada passo do algoritmo para aceitar a troca de posições dos nós. Para esta operação não se faz necessário o cálculo da função custo total, visto que será feita uma subtração entre o custo resultante e o inicial e todo ε_i que não se alterou será ignorado. Para esta operação basta calcular a soma dos custos que envolvem os nós selecionados para a troca e seus vizinhos, antes e depois da troca; com esta modificação, reduzimos o número de cálculos de ε_i de $2N^2$ para $12N$. É preciso ter cuidado para não calcular mais de uma vez o custo local de um mesmo par de nós.

A segunda mudança é uma questão de representação. A matriz de adjacência, que colaborou para a idealização do método e que é uma forma fácil de visualizar o ordenamento, não está incluída na forma mais barata de se calcular o ordenamento. O método que parte da matriz de adjacência e vai trocando linhas e colunas para representar a troca de ordem dos nós faz com que sejam realizadas $4N-4$ operações de troca.

Proponho um método análogo em que são utilizados três objetos, uma matriz e dois vetores. Estes são:

- **Matriz de interação**: matriz quadrada $N \times N$, simétrica, binária e com diagonal nula. Cada elemento, $I_{m,n}$, representa a existência ou não de ligação entre os nós indexados como m e n, onde 0 representa a ausência de interação e 1 a existência desta. Esta matriz não é alterada durante a execução do programa.

- **Vetor de posição**: vetor de tamanho N. Cada elemento do vetor, P_m, representa a posição do nó de índice m no ordenamento. Seus valores vão de 1 a N. Por exemplo, a proteína ENSP00000000412, indexada como 1, está na vigésima posição do ordenamento; então P_1 contém o valor 20.

- **Vetor de localização**: vetor de tamanho N. Cada elemento do vetor, L_i, representa o índice do nó que se encontra na posição i da lista. É a operação inversa ao vetor de

posição. Utilizando o mesmo exemplo anterior, L_{20} possui o valor 1. Fica claro que $L_{P_m} = m$ e $P_{L_i} = i$

O método de ordenamento utilizando estes vetores consiste em manter a matriz de interação intacta e fazer as alterações do ordenamento nos vetores de localização e posição. Ao invés das $4N-4$ operações de troca da matriz de adjacência do MFC, fazemos apenas 4 alterações. Para obter a representação da matriz de adjacência final, basta montar a matriz com a ordem obtida no vetor de localização. A modificação proposta causa aumento na quantidade de memória ocupada; eram utilizados $N \times N$ inteiros, agora são utilizados $N \times (N+2)$.

O programa foi compilado utilizando a opção -fast do compilador icc da Intel; esta opção habilita as otimizações -ipo, -O3, -no-prec-div, -static e -xHost. Estas opções tornam o programa um pouco mais pesado, utilizando mais memória RAM, e reduzem a sua precisão para cálculos envolvendo ponto flutuante, principalmente de divisão. Isto não é um problema para este método porque a maioria das variáveis de ponto flutuante são números aleatórios, não havendo necessidade de extrema precisão. O programa foi executado em computadores com processador Intel Core i7 870, com velocidade de *clock* de $2,9MHz$, $8MB$ de cache e $8MB$ de memória RAM.

Com a implementação destas propostas, o tempo de execução do método foi muito reduzido para as redes naturais, que possuem $N > 4.10^3$. O método que levaria um mês para realizar $3.10^3 MCS$ agora é resolvido em pouco mais de uma hora com $10^5 MCS$. Além disto, foram estas mudanças que indicaram o caminho seguido na implementação do método em mais de uma dimensão.

3.3 Ordenamento em duas dimensões

O grande problema que foi observado com clareza nas redes artificiais ordenadas em uma dimensão é a frustração da rede. A frustração também está presente nas redes naturais, cada ponto distante da diagonal da matriz de adjacência é um caso deste problema. Como pode ser observado, é impossível resolvê-lo com ordenamentos em uma dimensão. Proponho aqui um método de ordenamento em duas dimensões para verificar a redução das frustrações da rede segundo a quantidade de dimensões.

Escolhemos ordenar em duas dimensões pois ainda é possível a visualização das funções que queremos analisar, como a localização de ontologias, rotas metabólicas ou intensidade da expressão gênica. Agora, como o ordenamento é realizado em duas dimensões, a terceira di-

mensão é necessária para a representação das funções analisadas; para isso utilizamos a representação em escala de cores.

A partir do desenvolvimento do método com duas dimensões podemos generalizar o algoritmo para quantas dimensões desejarmos. O problema das distribuições em mais de duas dimensões é a impossibilidade da visualização dos resultados, tanto do ordenamento quanto da distribuição das funções estudadas.

Observamos que, somente depois da otimização do método proposta na seção 3.2, foi possível iniciar o desenvolvimento do método em duas dimensões. Isso ocorre porque a representação da matriz de adjacência e a implementação do método utilizando esta se tornam impossíveis para o ordenamento em duas dimensões. A matriz torna-se um tensor de quarta ordem, com $N \times N \times N \times N$, não visualizável e computacionalmente inviável para grandes valores de N.

As mudanças na implementação do método em $2D$ se dão nos vetores de localização de posição descritos na seção 3.2. A distribuição dos nós na matriz de localização é um problema, pois se N não possui raiz inteira, sobram elementos na matriz. Estes elementos nos obrigaram a considerar espaços vazios na matriz. A solução foi considerá-los como nós que não possuem nenhuma ligação. Com o surgimento destes espaços, o método fica um pouco modificado.

Primeiro indexamos os nós com valores inteiros não repetidos entre 1 e N. Montamos a matriz de interação e sorteamos as posições para os nós. As matrizes utilizadas no ordenamento em duas dimensões são descritas a seguir.

- **Matriz de interação**: matriz quadrada $(N+1) \times (N+1)$, simétrica e binária. Cada elemento, $I_{m,n}$, representa a existência ou não de ligação entre os nós indexados como m e n, onde 0 representa a ausência de interação e 1 a existência desta. Todos os elementos da diagonal ou com índice 0 são nulos, ou seja, $I_{m,m} = I_{0,m} = I_{m,0} = 0$.

- **Matriz de localização**: matriz de tamanho $S \times S$, onde $S \geq \sqrt{N}$. Cada elemento da matriz, $L_{i,j}$, representa o índice do nó que se encontra nas posição i e j segundo os eixos x e y do ordenamento. $L_{i,j}$ pode assumir valores inteiros entre 0 e N, onde 0 representa os espaços vazios.

- **Matriz de posição**: matriz de tamanho $(N+1) \times 2$. Cada elemento da matriz, $P_{m,i}$, representa a posição do nó m segundo o eixo i no ordenamento. Seus valores vão de 0 a S.

A função custo também precisa ser modificada. A função proposta por Rybarczyk *et al.*[12–14], descrita pela equação 2.9, mostra que esta possui dois fatores, o primeiro é o quadrado

da distância do elemento para a diagonal, o segundo confere se existem interfaces na vizinhança do elemento. É importante lembrar que esta equação foi proposta pensando na representação via matriz de adjacência e que somente pares ligados contribuem para o custo. Podemos ver que o primeiro termo se refere, na verdade, à distância entre os vértices do par ligante; já o segundo termo calcula a quantidade de vizinhos de um nó que não estão ligados ao outro vértice do par. Em duas dimensões, a função custo deve conter os mesmos fatores, tomando a forma das equações a seguir:

$$\varepsilon = \sum_{i,j=1}^{N} \varepsilon_{i,j} = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}|i-j|^{\alpha}(4 - a_{i,j+1} - a_{i,j-1} - a_{i+1,j} - a_{i-1,j}) \tag{2.9}$$

$$\varepsilon = \sum_{m,n=1}^{N} \varepsilon_{m,n} = \sum_{m,n=1}^{N} I_{m,n} D_{m,n}^{\alpha} F_{m,n} \quad , \tag{3.1}$$

onde $D_{m,n}$ é a distância entre os nós m e n e $F_{m,n}$ é a soma quantidade de vizinhos de n que não estão associados a m e vice-versa. Para duas dimensões, as componentes da equação 3.1 são

$$D_{m,n} = [(P_{m,0} - P_{n,0})^2 + (P_{m,1} - P_{n,1})^2]^{\frac{1}{2}}, \tag{3.2}$$

$$F_{m,n} = 8 - I_{m,n_1} - I_{m,n_2} - I_{m,n_3} - I_{m,n_4} - I_{m_1,n} - I_{m_2,n} - I_{m_3,n} - I_{m_4,n} \tag{3.3}$$

e onde m_i e n_i são os primeiros vizinhos de m e n, isto é,

$$\begin{aligned} m_1 &= L_{(P_{m,0}),(P_{m,1}+1)} \\ m_2 &= L_{(P_{m,0}+1),(P_{m,1})} \\ m_3 &= L_{(P_{m,0}),(P_{m,1}-1)} \\ m_4 &= L_{(P_{m,0}-1),(P_{m,1})} \end{aligned} \tag{3.4}$$

sendo similar para n.

O algoritmo do ordenamento em duas dimensões é muito semelhante ao de uma dimensão. Sorteamos índices para os nós e montamos a matriz de interação. Sorteamos posições para os vértices, montando as matrizes de posição e localização. Calculamos o custo total inicial e escolhemos um valor alto para o parâmetro de temperatura, como um milésimo do custo inicial.

Cada passo do ordenamento consiste em sortear um vértice e uma posição, calcular a soma de todos os custos locais que envolvam estes pontos, realizar a troca de posição e recalcular a soma dos custos. Se o custo diminuir aceitamos a alteração; caso este aumente, aceitamos a troca com probabilidade $P = e^{(-\frac{\Delta\varepsilon}{T})}$, onde $\Delta\varepsilon$ é a variação do custo e T é a temperatura. Mantemos este parâmetro constante por um grande número de passos de Monte Carlo e depois

diminuimos a temperatura.

3.3.1 Redes artificiais

Ordenamos em duas dimensões as redes artificiais simples criadas na seção 3.1.1, com as matrizes de interação das figuras 3.1 e 3.2. Os resultados obtidos estão descritos nas figuras 3.6 e 3.7. É possível observar que para a primeira rede o resultado se mantém ótimo, e para a segunda, ele apresenta uma melhor resolução dos módulos intermediários. Além da diminuição da frustração do resultado final, é esperado que, por haver menos frustrações e mais possibilidades de trocas, os mínimos locais da função de custo sejam menos estáveis, facilitando a atuação do *annealing*.

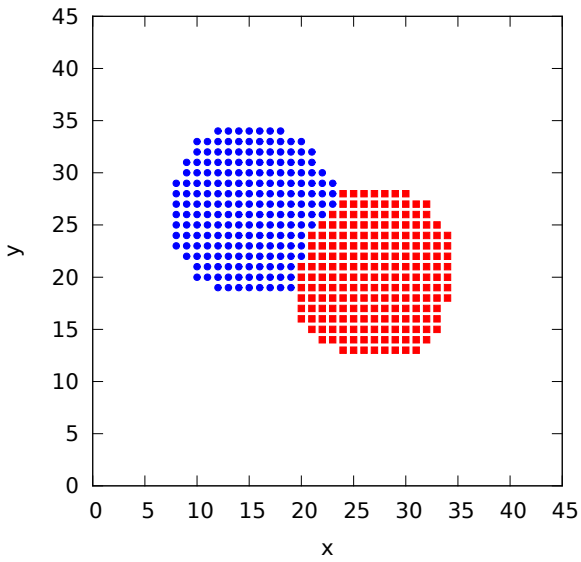

Figura 3.6: Ordenamento final da rede com interações descritas na tabela 3.1 e na figura 3.1, onde os pontos vermelhos pertencem ao módulo I e os azuis ao módulo II.

3.3.2 Redes naturais

Ordenamos em duas dimensões os proteomas de *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae*; as posições resultantes são semelhantes à figura 3.8. Estes resultados em que são exibidas apenas as posições não nos dão nenhuma informação relevante, e, além disso não podemos representar o resultado final em uma matriz de adjacência com quatro dimensões.

Podemos verificar a distribuição das associações em relação à distância entre os nós. Como pode ser observado na figura 3.9, existe uma relação direta entre essas características; quanto

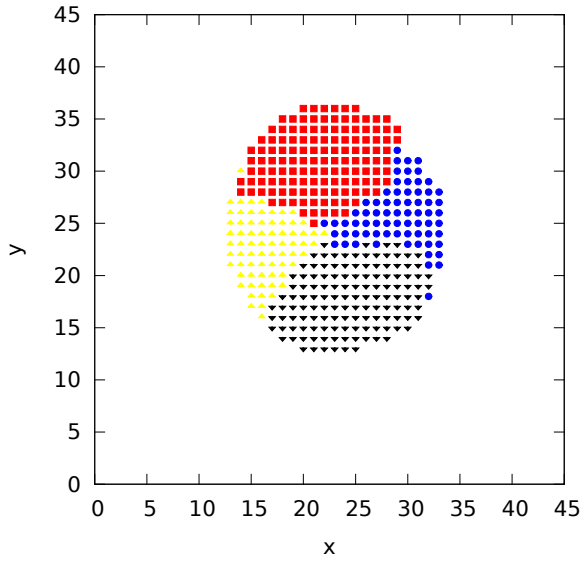

Figura 3.7: Ordenamento final da rede com interações descritas na tabela 3.2 e na figura 3.2, onde os pontos vermelhos pertencem ao módulo I, os azuis ao II, os amarelos ao III e os pretos ao IV.

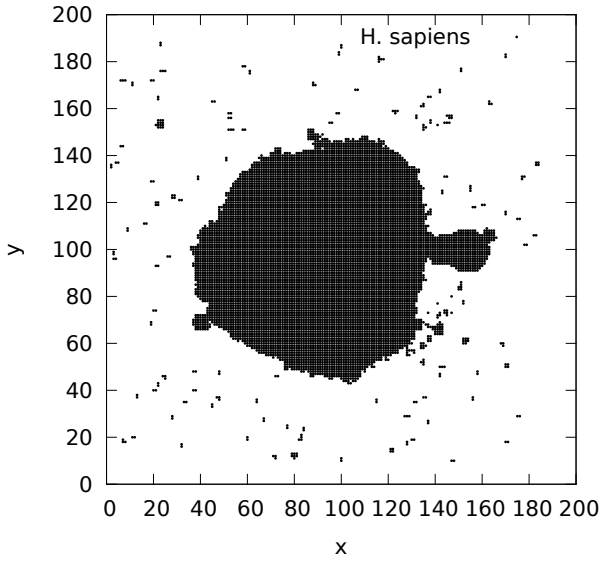

Figura 3.8: Ordenamento final da rede de proteínas de *Homo sapiens*.

mais próximas as proteínas estão, maior é a probabilidade de haver associação entre elas, como era esperado do ordenamento.

A análise do número de associações para cada distância entre proteínas associadas mostra que as redes de *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae* possuem características parecidas, como pode ser observado na figura 3.10. Para facilitar esta análise, calculamos a quantidade de associações cujas proteínas estão mais distantes que a

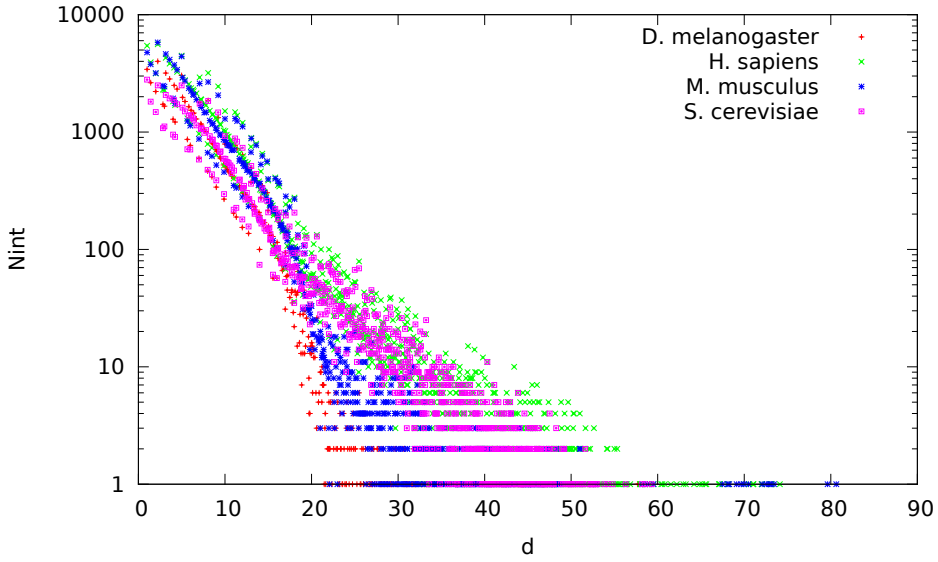

Figura 3.9: Número de associações por distância para os ordenamentos finais de *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae*.

distância d; denominamos este fator de $Cint$:

$$Cint(d) = 1 - \int_0^d \frac{I(x)}{Nint} dx, \tag{3.5}$$

onde $I(x)$ é a quantidade de associações cujas proteínas encontram-se à distância x e $Nint$ é o número total de associações. A figura 3.11 mostra claramente que, após o ordenamento da rede, a quantidade de associações cai exponencialmente com o aumento da distância entre proteínas.

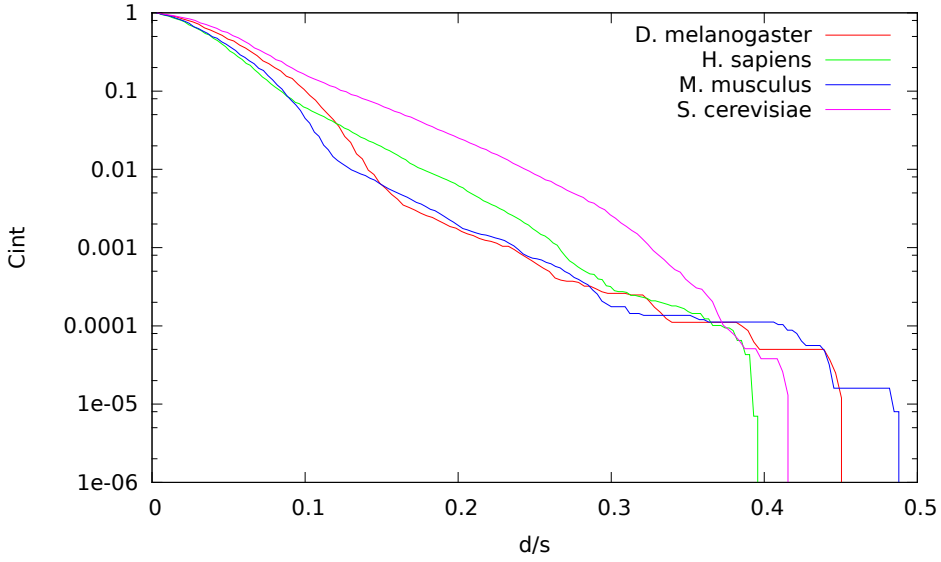

Figura 3.10: Quantidade de associações externas por distância, normalizada pelo tamanho da rede ordenada, para os ordenamentos finais de *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae* em duas dimensões.

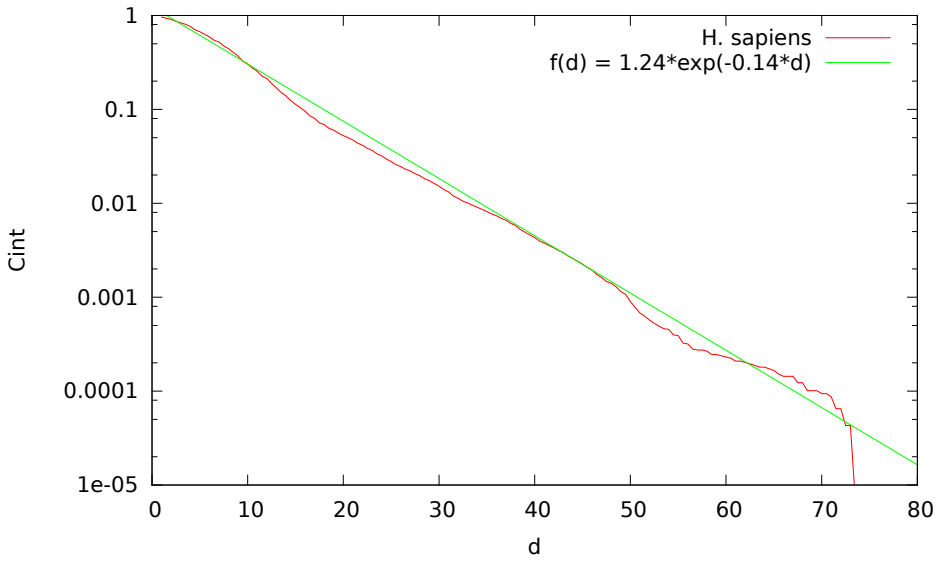

Figura 3.11: Quantidade de associações externas por distância para o ordenamento final de *Homo sapiens* em duas dimensões.

4 *Análise das redes*

Neste capítulo são discutidos os métodos empregados na análise da rede organizada em duas dimensões. Observamos a formação de módulos de associação das redes, fazemos a caracterização dos módulos funcionais para a rede de *Homo sapiens* e discutiremos o transcriptograma em duas dimensões, apresentando seu poder de diagnóstico.

4.1 Modularidade

A análise inicial do resultado trata de utilizar uma ferramenta chamada modularidade por janela. A idéia desta ferramenta é observar a razão entre a quantidade de ligações entre os nós da região R e o número de conexões total destes nós. Essa medida é realizada, segundo a equação 4.1, em todos os pontos que possuem algum nó ligante; a região selecionada é um círculo de raio r centrado no ponto escolhido. A modularidade assume valores entre 0 e 1, onde 0 significa que não há ligações entre os nós da região e 1 significa que todas as ligações dos vértices da região ocorrem dentro desta.

$$M_m = \frac{\sum_{i,j}^{R} I_{i,j}}{2(\sum_{i}^{R} k_i) - \sum_{i,j}^{R} I_{i,j}}. \tag{4.1}$$

A medida de modularidade por janela é eficaz para a observação de módulos com tamanho da ordem de πr^2. Módulos maiores apresentam modularidade baixa, pois em qualquer ponto do módulo, muitas ligações estarão fora da região selecionada para o cálculo. Módulos menores que a região também apresentarão baixa modularidade, visto que entre os nós selecionados para o cálculo, alguns participam de outros módulos. Por esses motivos, é muito importante ter cuidado ao interpretar a modularidade, que tanto depende da escolha do raio r.

As figuras 4.1 e 4.2, mostram os resultados do ordenamento de *Homo sapiens* através da medida de modularidade com raios 4 e 7, selecionando até 49 e 149 nós, respectivamente. O resultados dos outros organismos é análogo e concorda com a existência de módulos de

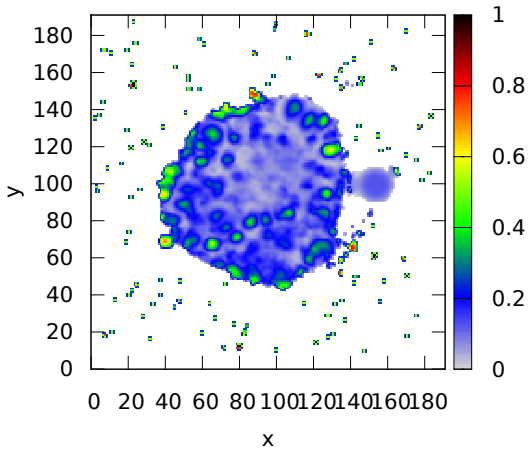

Figura 4.1: Modularidade com raio 4 do ordenamento final da rede de proteínas de *Homo sapiens*.

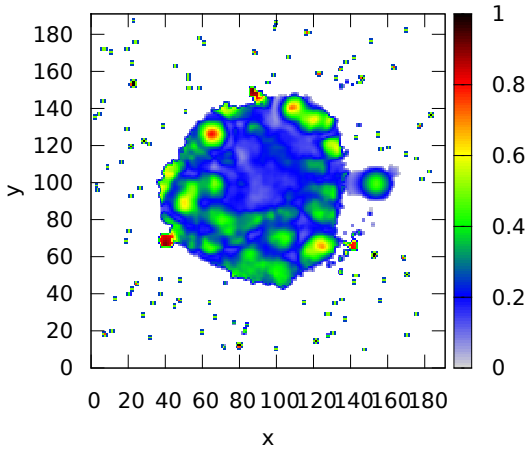

Figura 4.2: Modularidade com raio 7 do ordenamento final da rede de proteínas de *Homo sapiens*.

tamanhos variados, mostrando que o método de ordenamento bidimensional é coerente com o método MFC de uma dimensão.

4.2 Caracterização dos módulos

A caracterização funcional dos módulos de proteínas é um processo simples que se dá em poucos passos:

1. fazemos uma lista com as proteínas do módulo que desejamos analisar;

2. com a ferramenta de enriquecimento funcional AmiGO[37], citada na seção 2.3.2, descobrimos quais termos de ontologias são mais significativamente presentes nesta lista;

3. buscamos os termos de ontologias que nos interessam no sítio do *Gene Ontology*[36], citado na seção 2.3.2, fazendo uma lista com todas as proteínas que participam da função desejada;

4. para cada nó da rede ordenada, selecionamos as proteínas pertencentes a um círculo, com o raio desejado, centrado naquele nó;

5. calculamos a razão entre as proteínas selecionadas que participam desta função e o total de proteínas selecionadas e atribuímos este valor ao nó.

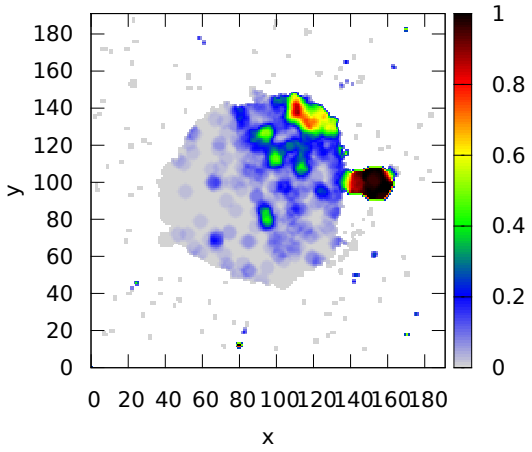

Figura 4.3: Distribuição de densidades do termo de ontologia GO:0004872, *Receptor activity*, calculada com raio 4, no ordenamento final da rede de proteínas de *Homo sapiens*.

O processo explicado nos fornece a função de distribuição de densidade daquela ontologia na rede. Exemplos de ontologias podem ser observadas nas figuras 4.3, 4.4, 4.5 e 4.6. Nessas figuras, fica claro o papel do ordenamento na caracterização do Proteoma.

4.3 Transcriptograma

O transcriptograma é uma técnica de análise de expressão gênica que trata de fazer médias dos dados de transcrição para aumentar a razão entre sinal e ruído[12–14]. Este método foi desenvolvido porque o ruído das medidas de expressão gênica por microarranjo é muito alto, tornando difícil resgatar informações precisas destes dados; supondo que proteínas relacionadas

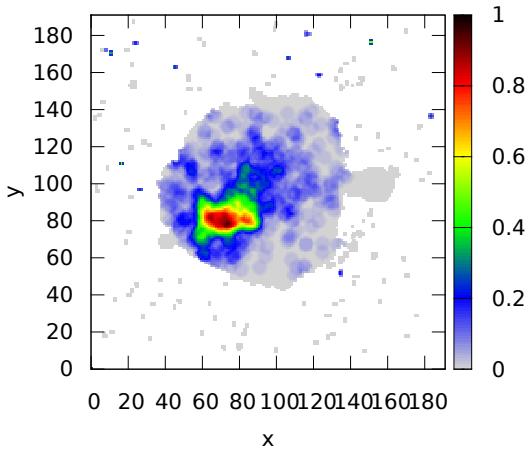

Figura 4.4: Distribuição de densidades do termo de ontologia GO:0007049, *Cell cycle*, calculada com raio 4, no ordenamento final da rede de proteínas de *Homo sapiens*.

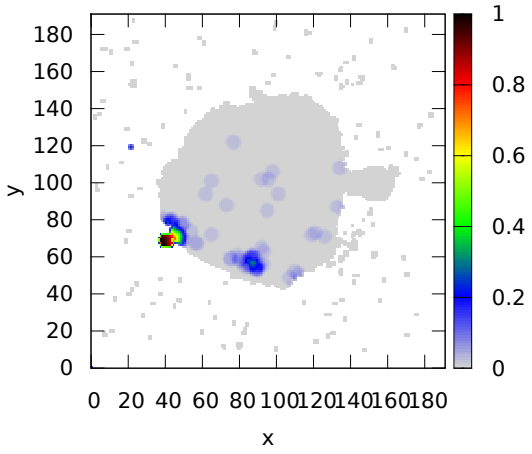

Figura 4.5: Distribuição de densidades do termo de ontologia GO:0045333, *Cellular respiration*, calculada com raio 4, no ordenamento final da rede de proteínas de *Homo sapiens*.

entre si se expressam em conjunto e que o ruído experimental é um ruído branco, uma média da expressão destas deve reduzir o erro e aumentar o sinal do grupo de proteínas.

Para calcular os valores da rede de transcriptograma em uma dimensão, selecionamos uma janela de tamanho definido, fazemos a média dos valores de transcrição das proteínas da janela e atribuímos este valor ao nó central da janela. Este processo é repetido para todos os nós da rede. O processo para fazer o transcriptograma em duas dimensões é análogo:

1. montamos a matriz M de transcrição, em que cada nó possui o valor de transcrição da

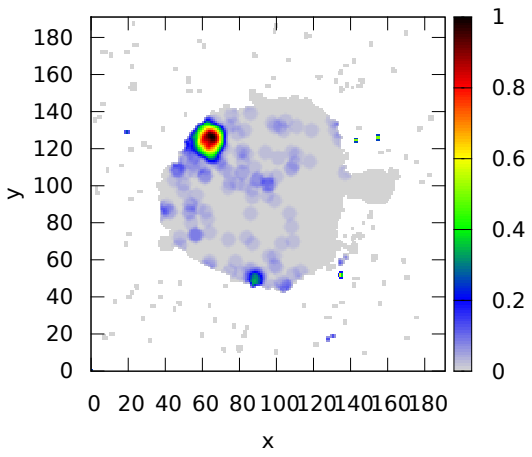

Figura 4.6: Distribuição de densidades do termo de ontologia GO:0006412, *Translation*, calculada com raio 4, no ordenamento final da rede de proteínas de *Homo sapiens*.

proteína respectiva a ele, ou é nulo, caso não tenha uma proteína relacionada;

2. para cada posição (i,j) onde há uma proteína, selecionamos todas as proteínas que se encontram a uma distância menor que o raio R da janela escolhida;

3. tomamos a média do valor de transcrição das proteínas selecionadas e atribuímos o resultado à posição (i,j) da matriz de transcriptograma;

4. para todas as posições onde não há proteínas, é atribuído o valor zero.

O resultado deste método pode ser descrito pela equação

$$t_{ij}^{(R)} = \frac{\sum_{kl}^{(\Delta x^2+\Delta y^2 \leq R^2)} M_{kl}}{\sum_{kl}^{(\Delta x^2+\Delta y^2 \leq R^2)} \theta(M_{kl})}, \tag{4.2}$$

onde $\Delta x = i-k$, $\Delta y = j-l$ e θ é a Função de Heaviside. Um exemplo de transcriptograma pode ser observado na figura 4.7, que mostra as médias locais de expressão gênica da amostra GSM337197 do experimento GSE13355[39, 40], que pode ser encontrado no banco de dados GEO. Esta amostra não apresenta sinal de psoríase.

4.3.1 Alterações funcionais causadas por psoríase

Psoríase é uma doença auto-imune que afeta a pele, manifestando-se através de inflamação, lesões avermelhadas descamativas e coceira. Como esta doença altera o estado metabólico das

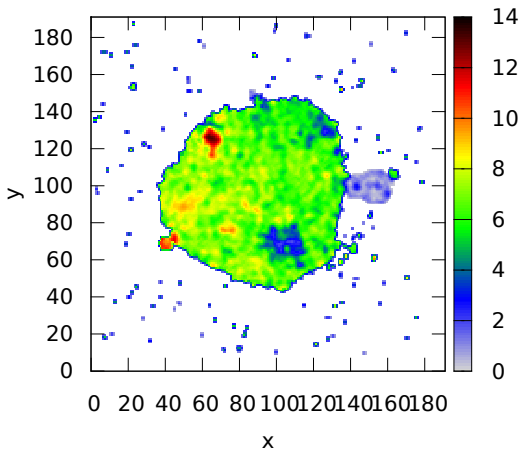

Figura 4.7: Transcriptograma do experimento GSE13355, amostra GSM337197, de *Homo sapiens* sem psoríase.

células afetadas, é esperado que os dados de expressão gênica mostrem alterações e, portanto, possam ser utilizados para realizar diagnósticos.

Para analisar as alterações da expressão gênica causada pela psoríase e encontrar os padrões de diagnóstico, usamos como base os experimentos de microarranjo GSE13355[39, 40] e GSE14905[41], encontrados no GEO. Estes experimentos, em conjunto, nos fornecem um total de 176 amostras, sendo 85 saudáveis e 91 doentes.

Calculamos o transcriptograma de todas as amostras disponíveis e então separamos estas em dois grupos: saudáveis e doentes. Calculamos as médias dos transcriptogramas de amostras saudáveis e doentes, $\langle t_{ij}^R \rangle_0$ e $\langle t_{ij}^R \rangle_1$, respectivamente; e desvios padrão, σ_0 e σ_1, formando um padrão de transcriptograma para cada grupo segundo as seguintes equações:

$$\langle t_{ij} \rangle_0 = \frac{1}{N_0} \sum_{\alpha=1}^{N_0} t_{ij}^{\alpha} \tag{4.3}$$

$$\langle t_{ij} \rangle_1 = \frac{1}{N_1} \sum_{\beta=1}^{N_1} t_{ij}^{\beta} \tag{4.4}$$

$$(\sigma_{ij}^2)_0 = \frac{1}{N_0} \sum_{\alpha=1}^{N_0} (t_{ij}^{\alpha} - \langle t_{ij} \rangle_0)^2 \tag{4.5}$$

$$(\sigma_{ij}^2)_1 = \frac{1}{N_1} \sum_{\beta=1}^{N_1} (t_{ij}^{\beta} - \langle t_{ij} \rangle_1)^2 \tag{4.6}$$

Os resultados podem ser observados nas figuras 4.8 e 4.9.

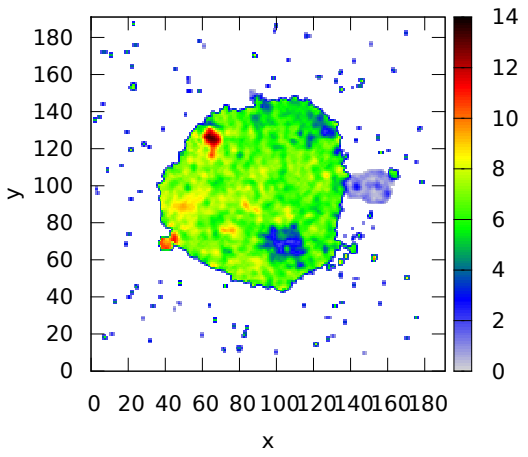

Figura 4.8: Média dos transcriptogramas de amostras sem psoríase.

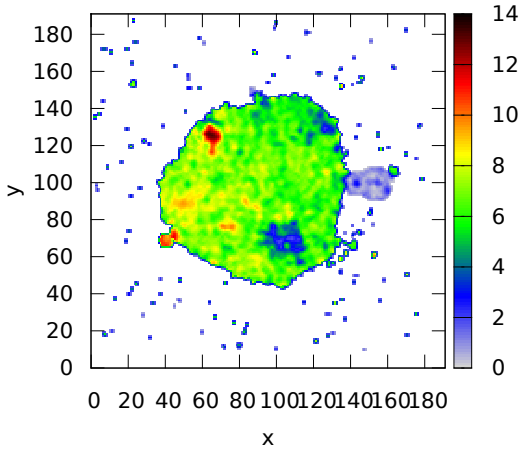

Figura 4.9: Média dos transcriptogramas de amostras com psoríase.

Observando as figuras 4.8 e 4.9, não é possível detectar modificações aparentes. Para observar estas alterações é necessário considerar o desvio padrão obtido para as amostras saudáveis, que é muito pequeno, figura 4.10. Como critério de comparação, normalizamos os resultados de forma a mostrar a alteração das médias em unidades de desvio padrão, $\langle t_{ij}^R \rangle_1^0$; para isso calculamos, ponto a ponto, a diferença entre o transcriptograma doente e o controle e dividimos pelo desvio padrão no ponto, como descrito na equação

$$\langle t_{ij}^R \rangle_1^0 = \frac{\langle t_{ij}^R \rangle_1 - \langle t_{ij}^R \rangle_0}{\sigma_0}. \tag{4.7}$$

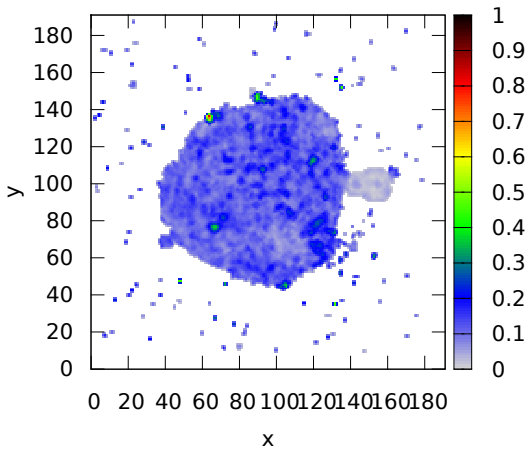

Figura 4.10: Desvio padrão dos transcriptogramas de amostras sem psoríase.

O resultado da normalização pode ser observado na figura 4.11. Neste gráfico podemos ver as alterações com precisão, e elas são bem aparentes em algumas regiões específicas. A região vinculada ao ciclo celular (GO:0007049), representada na figura 4.4, está bastante superexpressa, assim como outras duas regiões menores e sem funções biológicas específicas, mas existem variações mais suaves em quase todo o transcriptograma. É esperado que a função biológica de ciclo celular esteja superexpressa porque a inflamação das células causa aumento exagerado da proliferação destas, formando as placas avermelhadas características da doença.

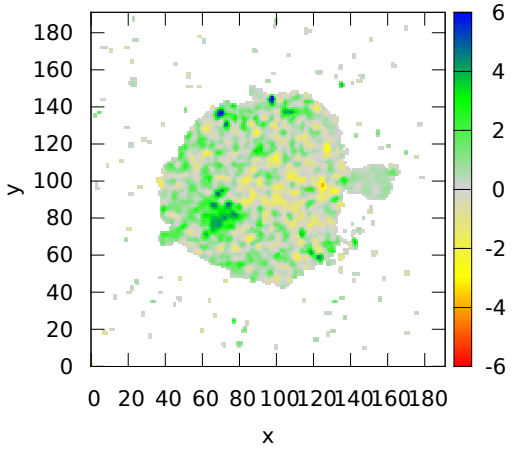

Figura 4.11: Diferença entre as médias de amostras doentes e saudáveis normalizada segundo as saudáveis, $\langle t_{ij}^{R} \rangle_{1}^{0}$.

4.3.2 Diagnóstico via transcriptograma

O método exposto na seção 4.3.1 nos permite ver com clareza as regiões alteradas pela doença e verificar, devido ao ordenamento, quais módulos funcionais estão subexpressos ou superexpressos. É possível, a partir dos padrões de transcriptogramas saudáveis e doentes, diagnosticar outras amostras, que sejam normalizadas segundo os padrões.

Classificamos os transcriptogramas em duas classes, saudáveis, 0, e doentes, 1; com N_0 e N_1 amostras consideradas padrão para cada classe. Para isso, seguimos os passos abaixo.

1. Produzimos um transcriptograma com raio R para cada amostra, ou seja, calculamos t_{ij}^{α} para a classe 0, t_{ij}^{β} para a classe 1 e t_{ij}^{γ} para a o grupo de teste; onde i e j são as posições no transcriptograma, $\alpha = 1, \ldots, N_0$, $\beta = 1, \ldots, N_1$ e $\gamma = 1, \ldots, N_t$.

2. Calculamos a média dos transcriptogramas para cada classe de controle, definidas pelas equações 4.3 e 4.4.

3. Calculamos o desvio padrão dos transcriptogramas para cada classe de controle segundo as equações 4.5 e 4.6.

4. Calculamos as distâncias relativas $\langle z \rangle$ de cada transcriptograma para as classes segundo as equações 4.8 e 4.9, onde N_{prot} é o número de posições do ordenamento ocupadas por proteínas.

$$\langle z \rangle_0 = \frac{1}{N_{prot}} \sum_{ij} \frac{(t_{ij} - \langle t_{ij} \rangle_0)^2}{(\sigma_{ij}^2)_0} \tag{4.8}$$

$$\langle z \rangle_1 = \frac{1}{N_{prot}} \sum_{ij} \frac{(t_{ij} - \langle t_{ij} \rangle_1)^2}{(\sigma_{ij}^2)_1} \tag{4.9}$$

5. Calculamos o índice de classificação x, descrito pela equação 4.10. Este índice pode ter valores entre 0 e 1.

$$x = \frac{\langle z \rangle_0}{\langle z \rangle_0 + \langle z \rangle_1} \tag{4.10}$$

6. Montamos a função de diagnóstico $D(x)$, que representa a probabilidade de uma amostra com índice de classificação x pertencer à classe 1. Para definir esta função, usamos as amostras de controle, α e β, cuja probabilidade de pertencer à classe 1 é conhecida, e fazemos o ajuste de curva para uma função sigmoide, descrita pela equação 4.11, onde A e B são os parâmetros de ajuste.

$$D(x) = \frac{1}{1 + e^{A - Bx}} \tag{4.11}$$

7. Para diagnosticar uma amostra γ, basta calcular o $x^{(\gamma)}$ e avaliar a probabilidade da amostra pertencer a cada classe do padrão segundo a função de diagnóstico. As probabilidades são calculadas segundo as equações 4.12 e 4.13.

$$P_0^{(\gamma)} = 1 - D(x^{(\gamma)}) \tag{4.12}$$

$$P_1^{(\gamma)} = D(x^{(\gamma)}) \tag{4.13}$$

O resultado do processo descrito, aplicado às amostras dos experimentos GSE13355 e GSE14905, é a função de diagnóstico descrita pela equação 4.14. Esta função pode ser observada no gráfico 4.12, juntamente com os dados de controle utilizados para o seu ajuste. É interessante ressaltar que a função $D(x)$ obtida possui uma faixa curta onde o diagnóstico é pouco preciso, mas a grande maioria das amostras encontra-se longe desta região e pode ser classificada com alto grau de precisão.

$$D(x) = \frac{1}{1 + e^{21,6463 - 45,0336x}} \tag{4.14}$$

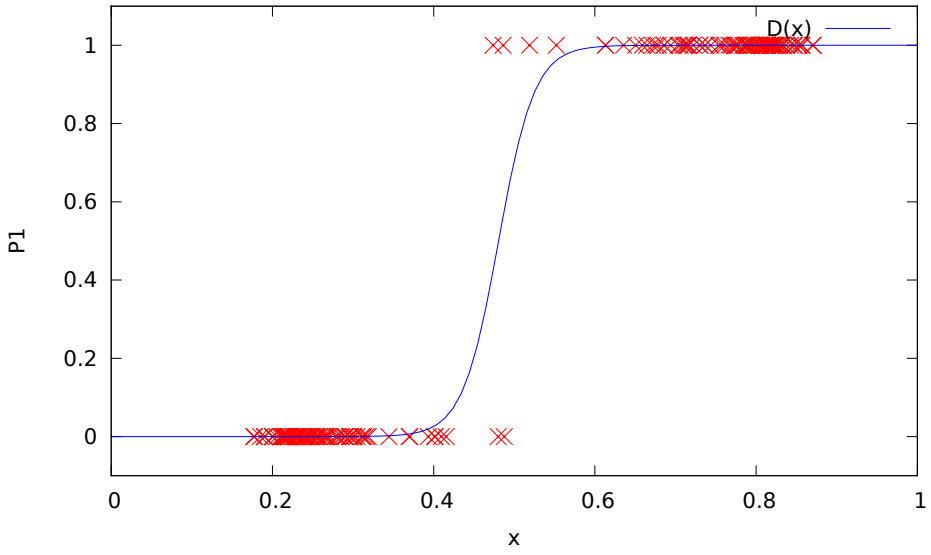

Figura 4.12: Função de diagnóstico, $D(x)$, e valores de x para as amostras de controle, onde $P1$ é a probabilidade de uma amostra pertencer à classe 1.

O mesmo processo foi realizado utilizando o ordenamento em uma dimensão, mas o resultado obtido não foi tão claro. A separação dos dois grupos de diagnóstico, saudável e doente, fica menos drástica, pois mais amostras encontram-se na região central. O diagnóstico, portanto, torna-se menos conclusivo quando realizado a partir de um ordenamento em uma dimensão, apesar de ainda ser interessante.

Com o método de diagnóstico a partir do transcriptograma em duas dimensões e seus resultados, participamos do desafio de diagnósticos do *Systems Biology Verification Improver*[42]. O desafio consistia em diagnosticar 62 transcriptomas feitos a partir de amostras de pele. O resultado deveria ser formatado na forma de uma matriz S, onde S_{mn} é a confiança de que a amostra m pertença à classe n e pode ter valores entre 0 e 1.

Entre as métricas de avaliação, a CCEM, *Correct Class Enrichment Metric*, estima o enriquecimento das classes avaliadas com correção. A CCEM é calculada de acordo com a equação 4.15, onde T_{mn} é 1 se a amostra m pertence à classe n, ou 0 no outro caso. Esta métrica soma os valores de confiança das amostras avaliadas corretamente e subtrai a confiança das avaliadas erroneamente, normalizada para que o resultado fique entre 0 e 1, sendo 1 o resultado máximo. Neste desafio, obtivemos valor de aproximadamente $0,9569$, segundo esta avaliação. Os nossos resultados nos renderam a décima quarta posição, entre 50 grupos participantes.

$$CCEM = \frac{\sum_m^N [\theta(S_{m0} - S_{m1}) T_{m0} S_{m0} - \theta(S_{m1} - S_{m0}) T_{m1} S_{m1}]}{2N} + 0,5 \tag{4.15}$$

5 *Conclusões*

5.1 Conclusões

Neste trabalho, apresentamos ajustes e otimizações para o método de ordenamento MFC. O resultado desta otimização foi uma redução no tempo de execução dos programas de um mês para uma hora. Esta otimização se deu a partir de mudanças conceituais necessárias para a implementação do método em duas dimensões, mas não alterou em nada o resultado final do ordenamento.

O novo método de ordenamento de redes em duas dimensões é um avanço em relação ao MFC, pois reduz as frustrações causadas por ordenar uma rede complexa em uma lista, além de abrir as portas para o ordenamento em N dimensões, facilmente adaptável a partir do método apresentado neste trabalho. A nova metodologia mostrou-se capaz de agregar os vértices da rede hierarquicamente, de forma que a probabilidade de haver ligação entre dois nós decai exponencialmente com o distancimento destes, como visto nas figuras 3.9, 3.10 e 3.11. A figura 5.1 deixa claro que, no ordenamento bidimensional, as proteínas associadas estão muito mais próximas do que no ordenamento unidimensional, reduzindo a frustração geométrica da rede.

A partir da rede já ordenada, podemos observar a estrutura de módulos e submódulos utilizando a ferramenta de modularidade por janela, como demonstrado nas figuras 4.1 e 4.2. É possível realizar o enriquecimento funcional, analisando o agrupamento de proteínas de uma mesma ontologia, observada nas figuras 4.3, 4.4, 4.5 e 4.6.

O transcriptograma bidimensional, gerado a partir do ordenamento e dos dados de transcriptoma, nos permite realizar o mapeamento da expressão do Proteoma analisado, visto na figura 4.7. Podemos verificar quais são as alterações das funções celulares, tanto subexpressão quanto superexpressão, causadas por alguma patologia ou fármaco, como na figura 4.11. O conhecimento sobre esses fatores pode ajudar a encontrar novos alvos para fármacos ou observar efeitos colaterais da aplicação destes.

A partir do transcriptograma em duas dimensões, também é possível fazer diagnósticos,

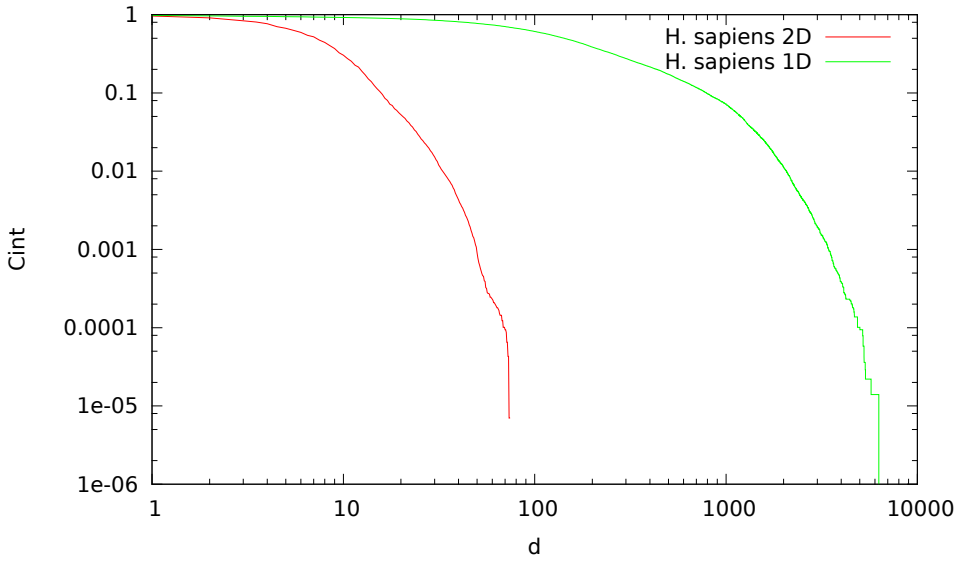

Figura 5.1: Distribuição de associações em relação à distância entre proteínas para os ordenamentos de *H. sapiens* em uma e duas dimensões

através da criação de padrões para estas, visto nas figuras 4.8 e 4.9 e aprimorado na figura 4.11. O método do transcriptograma melhora a razão entre sinal e ruído em relação à aplicação dos padrões diretamente no transcriptoma. A comparação dos diagnósticos em uma e duas dimensões mostra que os resultados do método bidimensional são mais precisos.

O ordenamento em duas dimensões, entretanto, tem problemas. A visualização de resultados é complicada, pois qualquer função $F(x,y)$ precisa ser representada em um terceiro eixo, como a escala de cores; isso impossibilita a visualização de mais de uma função em um mesmo gráfico, o que torna comparações entre resultados um tanto trabalhosa e complicada.

5.2 Perspectivas

O método de ordenamento mostrou-se promissor, apresentando bons resultados em uma dimensão e melhorias quando aprimorado para duas dimensões. Esta melhora deixa claro que este projeto não está finalizado e que ainda há muitos objetos de estudo nesta linha de pesquisa. O mesmo se aplica ao método de diagnóstico proposto.

Algumas das perspectivas tratam do aprimoramento do ordenamento, visando aumentar a razão sinal-ruído; outras tratam da análise dos resultados, visando melhorar a nossa interpretação dos resultados. É possível, aproveitando a experiência adquirida, criar um novo método, mais específico para diagnósticos. Algumas das idéias para futuros trabalhos são discutidas nesta seção.

5.2.1 Janela de avaliação

Todas as medidas de médias avaliadas neste trabalho foram realizadas utilizando raio de janela fixo e peso igual para todos os nós da janela. Considerando que a probabilidade de haver associação entre dois vértices decai com a distância, é válido imaginar que utilizar pesos diferentes para distâncias diferentes em relação ao centro da janela pode trazer resultados interessantes. Podemos também vincular a medida de uma posição ao vértice que nesta se encontra, é possível escolher o tamanho da janela de avaliação relacionado à conectividade do nó central. Será estudado e efeito destas diferentes formas de escolha da janela, de modo a melhorar os resultados da análise dos dados.

5.2.2 Ordenamento em N dimensões

Este trabalho mostra que o ordenamento bidimensional possui mais capacidade de organizar a rede do que o unidimensional, mas isso ocorre com o prejuizo da visualização dos resultados, que é dificultada e mais trabalhosa em duas dimensões. Com as propostas do capítulo 3, podemos estender o ordenamento para N dimensões sem dificuldades e, acredito eu, sem grande acréscimo de custo computacional. O aumento em dimensões do ordenamento pode trazer melhoras para o diagnóstico e avaliações da rede, mas ao custo total da visualização do ordenamento, podendo ser avaliado apenas matematicamente, mas sem maiores dificuldades.

5.2.3 Nova proposta de método para diagnóstico

A melhora dos resultados de diagnóstico devido ao aumento no número de dimensões do ordenamento ocorre por causa da diminuição das frustrações da rede ordenada. Tendo em vista que a quantidade de frustrações deve se manter elevada para qualquer número de dimensões muito menor que o tamanho da rede, talvez devessemos mudar o enfoque do método.

A idéia que surgiu durante a finalização deste trabalho é propor um método que não tenha o foco no ordenamento de redes e, muito menos, na visualização desta, mas apenas na aplicação desejada, como o diagnóstico com base em transcriptoma. É possível, conhecendo o Interatoma, desenvolver um método que aumente a razão entre sinal e ruído de um transcriptoma através de médias que envolvem a proteína avaliada e aquelas associadas a ela, direta ou indiretamente.

O método do transcriptograma consiste em calcular a média da expressão gênica sobre uma região da rede centrada ordenada para criar o transcriptograma, com valores referentes à posição das proteínas, t_{ij}. O novo método consistiria em calcular a média da expressão gênica sobre

uma proteína e todas as outras ligadas a esta por até N arestas da rede, com isso montaríamos o transcriptograma com valores referentes às proteínas, t_m. Com este método poderíamos montar um transcriptograma sem as frustrações causadas pela projeção de uma rede complexa sobre um sistema com poucas dimensões.

Este novo método pode se mostrar computacionalmente mais caro, dependendo do grau de associações a ser avaliado; outro problema seria a impossibilidade da visualização dos resultados do transcriptograma de forma gráfica, assim como com os ordenamentos para mais de duas dimensões. Estes custos são muito baixos quando comparados aos benefícios que podemos obter com o aumento da qualidade dos diagnósticos.

O diagnóstico seria, em princípio, realizado da mesma forma que neste trabalho. A partir dos dados do transcriptograma, seriam montados padrões para as classes avaliadas e verificadas a qual destas as amostras mais se assemelham.

5.2.4 Propostas de diferentes métricas para diagnósticos

O método de medida de distância de uma amostra para os padrões utilizada neste trabalho é muito simples e leva em conta todo o transcriptograma. É possível desenvolver uma métrica que considere apenas as regiões relevantes para o diagnóstico, excluindo informações que podem levar a erros no diagnóstico. É sabido que algumas regiões do transcriptograma caracterizam outras variáveis que não queremos avaliar, como o gênero do paciente; se excluímos esta porção e outras que não devem ser consideradas, podemos obter melhores resultados de diagnósticos.

Uma forma de minimizar este problema seria considerar para o diagnóstico apenas as regiões em que a distância entre os padrões seja maior que a soma dos desvios padrão obtidos para as classes. Outra métrica possível é valorizar as regiões em que os padrões mais se distanciem, dando peso maior para estas. Existem inúmeras possibilades para aprimorar o método proposto neste trabalho.

5.2.5 Disponibilização do método na rede

O método de MFC bidimensional se mostrou eficiente e deve ser disponibilizado para outros grupos. É parte do projeto desenvolver um programa capaz de realizar as tarefas propostas neste trabalho de forma fácil e simples, para isso devemos fazer um aplicativo multiplataforma com interface gráfica que integre os vários métodos envolvidos no ordenamento de redes, avaliação deste ordenamento, cálculo de transcriptograma e diagnóstico. Este aplicativo deve ser disponibilizado no sítio do grupo para *download*.

A disponibilização de um programa eficiente e simples, como o proposto acima, ajuda a divulgar o trabalho e incentiva o aprimoramento do método por outros grupos de trabalho, com outros enfoques e especialidades. Essa dinâmica seria muito proveitosa e poderia ampliar a abrangência do método MFC a níveis que seríamos incapazes de alcançar sozinhos.

Referências Bibliográficas

1 WATSON, J. D.; CRICK, F. H. Genetical implications of the structure of deoxyribonucleic acid. **Nature**, London, UK, v.171, n.4361, p.964-967, May, 1953.

2 WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. **Nature**, London, UK, v.171, n.4356, p.737-738, Apr, 1953.

3 CRAWFORD, M. H. HUGO: Genome data open to scientists. **Science**, Washington, USA, v.246, n.4937, p.1565, Dec, 1989.

4 OF HEALTH ENVIRONMENTAL RESEARCH, O. Human genome (1989-90 program report). Washington, USA: United States Department of Energy, 1990. Relatório técnico.

5 CHERRY, J. M. et al. SGD: Saccharomyces Genome Database. **Nucleic Acids Res**, Oxford, UK, v.26, n.1, p.73-79, Jan, 1998.

6 ASHBURNER, M.; DRYSDALE, R. Flybase - the drosophila genetic database. **Development**, Hampshire, USA, v.120, n.7, p.2077-2079, Jul, 1994.

7 SWARBRECK, D. et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. **Nucleic Acids Res**, Oxford, UK, v.36, n.SI, p.D1009-D1014, Jan, 2008.

8 KESELER, I.M. et al. EcoCyc: a comprehensive database of Escherichia coli biology. **Nucleic Acids Res**, Oxford, UK, v.39, n.Database issue, p.D583-D590, Jan, 2011.

9 NELSON, D. L.; COX, M. M. **Lehninger Principles of Biochemistry, Fourth Edition.** Fourth Edition. ed. New York, USA, 2004.

10 TOYODA, T. et al. Omicbrowse: a browser of multidimensional omics annotations. **Bioinformatics**, Oxford, UK, v.23, n.4, p.524-526, Feb, 2007.

11 GÓES-NETO, A. et al. Comparative protein analysis of the chitin metabolic pathway in extant organisms: A complex network approach. **Biosystems**, Oxford, UK, v.101, n.1, p.59-66, Jul, 2010.

12 RYBARCZYK-FILHO, J. L. et al. Towards a genome-wide transcriptogram: the Saccharomyces cerevisiae case. **Nucleic Acids Res**, Oxford, UK, v.39, n.8, p.3005-3016, Apr, 2011.

13 RYBARCZYK FILHO, J. L. **Medidas de performance metabólica usando a expressão gênica de genoma completo**. 2011. 112 f. Tese (Doutorado em Física) - Instituto de Física, Universidade Federal do Rio Grande do Sul, Porto Alegre. 2011.

14 BENETTI, F. **Homo sapiens: análise de expressão gênica por transcriptograma**. 2010. 35 f. Dissertação (Mestrado em Física) - Instituto de Física, Universidade Federal do Rio Grande do Sul, Porto Alegre. 2010.

15 GAUSE, W. C.; ADAMOVICZ, J. The use of the PCR to quantitate gene-expression. **PCR-methods and applications**, New York, USA, v.3, n.6, p.S123-S135, Jun, 1994.

16 SCHENA, M. et al. Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. **Science**, Washington, USA, v.270, n.5235, p.467-470, Oct, 1995.

17 WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, London, UK, v.10, n.1, p.57-63, Jan, 2009.

18 LOSCALZO, J.; KOHANE, I.; BARABASI, A.-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. **Mol Syst Biol**, New York, USA, v.3, n.124, Jul, 2007.

19 BARABÁSI, A.-L. Network medicine-from obesity to the "diseasome". **N Engl J Med**, Massachusetts, USA, v.357, n.4, p.404-407, Jul, 2007.

20 HIDALGO, C. A. et al. A dynamic network approach for the study of human phenotypes. **PLoS Comput Biol**, San Francisco, USA, v.5, n.4, p.e1000353, Apr, 2009.

21 GOH, K.-I. et al. The human disease network. **Proc Natl Acad Sci U S A**, Washington, USA, v.104, n.21, p.8685-8690, May, 2007.

22 NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. **The Structure and Dynamics of Networks: (Princeton Studies in Complexity).** Princeton, NJ, USA: Princeton University Press, 2006.

23 PAULO OSWALDO, B. N. **Grafos: teorias,modelos e algoritmos.** 4ª. ed. São Paulo, Brasil: Edgar Blücher, 2001. 328p.

24 GERSTING, J. L. **Fundamentos matemáticos para ciência da computação.** 3ª. ed. São Paulo, Brasil: LTC, 1995.

25 RAVASZ, E. et al. Hierarchical organization of modularity in metabolic networks. **Science**, Washington, USA, v.297, n.5586, p.1551-1555, Aug, 2002.

26 RAVASZ, E.; BARABÁSI, A.-L. Hierarchical organization in complex networks. **Phys Rev E Stat Nonlin Soft Matter Phys**, New York, USA, v.67, n.2 Pt 2, p.026112, Feb, 2003.

27 ERDÖS, P.; RÉNYI, A. On the evolution of random graphs. In: PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, 1960. [s.n.], 1960. p.17-61.

28 CAMIN, J. H.; SOKAL, R. R. A method for deducing branching sequences in phylogeny. **Evolution**, Malden, MA, USA, v.19(3), p.311-326, 1965.

29 BARABÁSI, A.-L.; BONABEAU, E. Scale-free networks. **Sci Am**, London, UK, v.288, n.5, p.60-69, May, 2003.

30 BARABÁSI, A.-L.; RAVASZ, E.; VICSEK, T. Deterministic scale-free networks. **Physica A: Statistical Mechanics and its Applications**, New York, USA, v.299, n.3-4, p.559-564, 2001.

31 VON MERING, C. et al. String: known and predicted protein-protein associations, integrated and transferred across organisms. **Nucleic Acids Res**, Oxford, UK, v.33, n.Database issue, p.D433-D437, Jan, 2005.

32 VON MERING, C. et al. String 7-recent developments in the integration and prediction of protein interactions. **Nucleic Acids Res**, Oxford, UK, v.35, n.Database issue, p.D358-D362, Jan, 2007.

33 JENSEN, L. J. et al. String 8-a global view on proteins and their functional interactions in 630 organisms. **Nucleic Acids Res**, Oxford, UK, v.37, n.Database issue, p.D412-D416, Jan, 2009.

34 SZKLARCZYK, D. et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. **Nucleic Acids Res**, Oxford, UK, v.39, n.Database issue, p.D561-D568, Jan, 2011.

35 KANEHISA, M. et al. KEGG for integration and interpretation of large-scale molecular data sets. **Nucleic Acids Res**, Oxford, UK, v.40, n.Database issue, p.D109-D114, Jan, 2012.

36 ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. **Nat Genet**, London, UK, v.25, n.1, p.25-29, May, 2000.

37 CARBON, S. et al. AmiGO: online access to ontology and annotation data. **Bioinformatics**, Oxford, UK, v.25, n.2, p.288-289, Jan, 2009.

38 EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. **Nucleic Acids Res**, Oxford, UK, v.30, n.1, p.207-210, Jan, 2002.

39 NAIR, R. P. et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappa B pathways. **Nature Genetics**, New York, USA, v.41, n.2, p.199-204, Feb, 2009.

40 SWINDELL, W. R. et al. Genome-Wide Expression Profiling of Five Mouse Models Identifies Similarities and Differences with Human Psoriasis. **Plos One**, San Francisco, USA, v.6, n.4, Apr, 2011.

41 YAO, Y. H. et al. Type I Interferon: Potential Therapeutic Target for Psoriasis?. **Plos One**, San Francisco, USA, v.3, n.7, Jul, 2008.

42 MEYER, P. et al. Industrial methodology for process verification in research (IMPROVER): towards systems biology verification. **Bioinformatics**, Oxford, UK v.28, n.9, p.1193-1201, May, 2012.