UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

Medidas de performance metabólica usando a expressão gênica de genoma completo

José Luiz Rybarczyk Filho

Tese de Doutorado
Porto Alegre
2011

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA
Tese de Doutorado

Medidas de performance metabólica usando a expressão gênica de genoma completo *

José Luiz Rybarczyk Filho

Tese de Doutorado, realizado sob orientação da professora Dra. Rita Maria Cunha de Almeida, apresentada ao Instituto de Física da UFRGS em preenchimento dos requisitos para a obtenção do título de Doutor em Ciências.

Porto Alegre
2011

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

Dedico

A Eleonora, minha mãe

Agradecimentos

- À Professora Rita M. C. de Almeida pela orientação dedicada, pelo incetivo e pela amizade.

- Aos Professores Leonardo G. Brunnet (IF-UFRGS) e José Cláudio F. Moreira (Bioquímica-UFRGS) pela colaboração durante a realização desta Tese.

- Aos amigos de grupo de pesquisa, Mauro A. A. Castro, Rodrigo J. S Dalmolin, Fernanda Benetti e Ricardo F. Mello.

- Aos meus amigos da sala M203, Samir, Daniel, Gustavo, Ana Paula, Felipes Bregolin. E aos amigos externos a sala M203, Ana Cláudia, Felipe Kremer e Cecília.

- À minha mãe Eleonora e ao meu irmão Jean Carlos, que sempre me apoiaram em todos os momentos da minha vida.

- Aos meus avós Irma *(in Memoriam)* e Oswaldo.

- Enfim, agradeço a todos que de alguma forma contribuíram para elaboração desta Tese.

Resumo

A cada dia surgem novas tecnologias que possibilitam aos cientistas medirem a expressão de inúmeros genes em uma única experiência com uma alta rapidez e eficiência. No entanto, ainda não existem muitas metodologias que sirvam para a análise do funcionamento de células e tecidos que possibilitem diagnóstico, prevenção e terapias de doenças. Na presente tese propomos uma metodologia de análise de expressão gênica que mostra diferenças na performance da célula com o passar do tempo, e tem poder de diagnóstico quando uma célula mutada é comparada com uma célula normal. Partimos da idéia de que rotas bioquímicas podem ser representadas por uma rede de interações entre metabólitos, entre os quais muitos deles são proteínas codificadas a partir do genoma. Sendo assim, o funcionamento de uma célula implica em interações que compõem uma rede intricada e complexa.

Reorganizamos a lista de genes em uma dimensão e reescrevemos a matriz de interação, sem a criação ou a destruição de interações, buscando correlacionar cada um dos genes com os seus respectivos vizinhos na lista unidimensional. Logo após a reorganização, encontramos módulos funcionais sobre a lista ordenada que são independentes do estado da célula estudada, é possível reconhecer os processos biológicos de cada módulo com o uso de banco de dados específicos.

Com base nisto, sobrepondo dados de expressão gênica sobre o ordenamento (transcriptograma), teremos um perfil transcricional de um tecido no ato da medida experimental. Se medirmos a expressão gênica de células provenientes do mesmo tecido em diferentes estágios do ciclo celular podemos seguir as alterações metabólicas ao longo do ciclo celular. Também poderíamos analisar a expressão gênica de um tecido normal com um tecido alterado. Como resultado desta comparação saberíamos quais módulos estão super expressos e os subexpressos em relação ao tecido normal. Publicamos na internet um software que é capaz de reproduzir essas análises e gerar transcriptogramas para análise de expressão gênica.

Abstract

Every day new technologies which allow scientists to measure the expression of numerous genes in a single experiment with a high speed and efficiency are emerging. However, there aren't many methods which are used to analyze the functioning of cells and tissues which allow diagnosis, prevention and therapy of diseases. In this thesis we propose a methodology for gene expression analysis that presents those differences in cell performance over time, and has diagnostic power when comparing a mutant cell with a normal one. Starting from the idea that biochemical pathways can be represented by a network of interactions between metabolites, among which many are encoded proteins from the genome. Thus, the functioning of a cell involves interactions that make up an intricate and complex network.

We reorganize the genes list in one dimension and rewrite the matrix of interaction, without the creation or destruction of interactions, seeking to correlate each gene with their respective neighbors in a one-dimensional list. Soon after the reorganization, we find functional modules on the ranked list that are state independent of the cell studied, it is possible to recognize the biological processes for each module by using specific databases.

Based on this, overlapping gene expression data on reorganized gene list (transcriptogram), we obtain a transcriptional profile of a tissue at the experimental measurement time. If we measure the gene expression of cells from the same tissue at different celluar stages we can follow the metabolic changes during the cell cycle. We could also analyze the gene expression of normal tissue with a mutant tissue. As a result this comparison we would know what modules are super expressed and underexpressed when compared to normal tissue. We published on the Internet software which is able to reproduce these analysis and generate transcriptograms for gene expression analysis.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Revisão

1.1 Introdução

Em 1953 James Watson e Francis Crick anunciaram ao mundo que haviam descoberto o segredo da vida [WAT 53a, WAT 53b], desvendando a estrutura do DNA (ácido desoxirribonucléico) e, em 1962, ganharam o prêmio Nobel em medicina e fisiologia por tal descoberta. Houve uma revolução na ciência, pois o DNA contém todas as informações genéticas de um organismo e está presente em cada célula que compõe tal ser vivo.

Em 1990, o Departamento de Energia (DOE) e o Instituto Nacional de Saúde (NIH) dos Estados Unidos da América iniciaram o Projeto Genoma com um financiamento de 50 bilhões de dólares e com duração prevista de 15 anos. O Projeto é um trabalho conjunto de vários países visando sequenciar o código genético de organismos vivos, sejam eles vegetais, animais, bactérias, fungos, archae, vírus, etc. O objetivo principal era criar mapas físicos de alta resolução, sequenciar todo o DNA do genoma Humano e criar bancos de dados com essa informação. Em 1998 os Estados Unidos, Japão, Austrália e alguns países da Europa se uniram para formar um organização internacional para coordenar o projeto de sequenciamento do *Homo sapiens*, HUGO (Human Genome Organisation) [CRA 89, oHER 90]. A função do HUGO é então organizar o trabalho de sequenciamento do genoma humano, armazenando o conhecimento obtido em um banco de dados centralizado, que é chamado de Genome Database, analisar o genoma do ponto de vista funcional e promover a aplicação destes conhecimentos para o melhoramento da saúde humana. Em 2003 o mapa físico foi concluído e o HUGO tem desempenhado um novo papel: a disseminação de dados funcio-

nais do genoma e o fornecimento das diretrizes responsáveis para as aplicações e implicações do genoma. Existem outras organizações equivalente ao HUGO, mas para outros organismos: Saccharomyces Genome Database (SGD) para *Saccharomyces cerevisiae* [CHE 98], EcoCyc para *Escherichia coli*, Flybase para *Drosophila melanogaster* [ASH 94], TAIR para *Arabidopsis thaliana* [SWA 08], etc.

O sequenciamento do DNA, porém, não é suficiente para uma compreensão do funcionamento celular. Faz-se necessário entender também os processos de transcrição do DNA, isto é, os processos pelos quais a informação contida no DNA são traduzidas em proteínas que, de uma maneira geral, são as moléculas que levam a cabo as reações bioquímicas no interior da célula. A seguir, é preciso descrever como essas reações podem dar lugar ao metabolismo celular, isto é, à vida celular. No processo de transcrição, primeiramente é produzido RNA a partir de regiões relativamente pequenas do DNA. O RNA é uma molécula similar ao DNA, com as importantes diferenças de ser uma molécula de uma única hélice (em contraposição ao DNA que é uma dupla hélice) e, no caso de transcrição, muito mais curta que o DNA. Essas características possibilitam que o RNA transcrito seja transportado até o citoplasma da célula onde, com uso intensivo de energia são produzidas as proteínas. Essa produção respeita a informação contida no DNA: cada uma das bases do DNA (adenina, citosina, guanina, timina) tem a sua complementar no RNA transcrito (adenina, citosina, guanina, uracila), que por sua vez é traduzida em componente protéico. Cada 3 bases dão lugar a um aminoácido e um conjunto de aminoácidos formam uma proteína. As regiões do DNA que codificam uma proteína são chamadas de genes [NEL 04].

Na figura 1.1, apresentamos algunas componentes de uma célula e através dele podemos ver uma complexidade química e estrutural cada vez maior. Para cada uma das etapas desse fluxograma estão sendo desenvolvidas abordagens metodológicas que visam desvendar os segredos da célula. Assim, o **genoma** considera a molécula de DNA e as informações armazenada sob a forma de genes, íntrons, exons, etc. A transcrição do DNA para o RNA, a primeira etapa do fluxo de informação genética, é investigada pelo estudo do **transcriptoma**. Nem todas as partes do DNA podem ser traduzidas em RNA; uma fração importante do DNA nunca é transcrito e, portanto, não dá lugar a proteínas. O papel destas partes no DNA ainda é alvo de investigação e as hipóteses levantadas consideram vantagens estrutu-

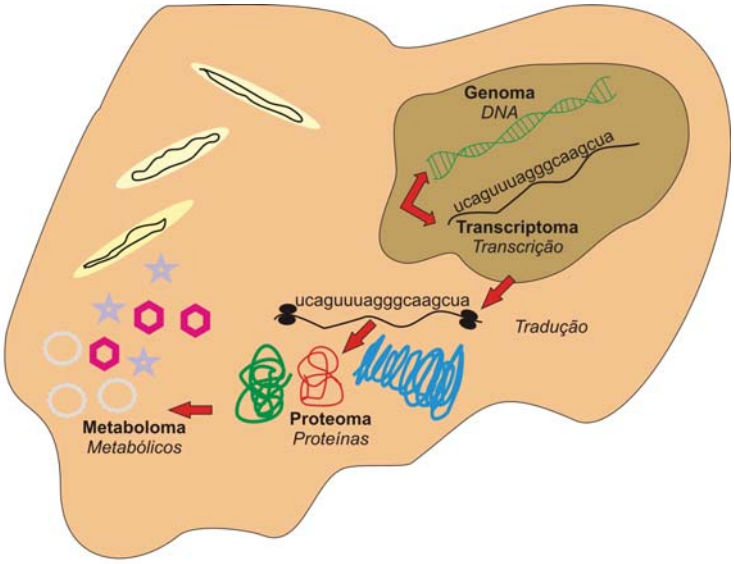


Fig. 1.1: *Fluxo de informação genética em uma célula: a informação no DNA é repassada para o RNA, que a leva até os centros em que as proteínas são produzidas. Essas são as principais moléculas envolvidas nas cascatas bioquímicas no interior da célula, embora também moléculas ou íons mais simples (como $Ca^{+}2$) também desempenhem papéis importantes.*

rais, que possibilitariam ao DNA configurações de dobramento estáveis, ou seriam resquícios da evolução sofrida pelo DNA desde a primeira célula. Células de um mesmo organismo, mas originárias de diferentes tecidos possuem o mesmo DNA, mas expressam-no diferentemente. Consequentemente o transcriptoma das células pode variar dependendo da sua diferenciação e do meio ambiente. Assim, o transcriptoma fornece informação a respeito de quais genes estão sendo expressos em um determinado momento. O **proteoma**, por outro lado, detecta diretamente as proteínas expressas pelas células. A diferença entre o transcriptoma e o proteoma consiste nas técnicas experimentais que acessam RNA antes da sua tradução em proteínas, ou já diretamente as proteínas. Já o **metaboloma** considera todos os componentes que podem participar das reações bioquímicas no interior da célula, incluido assim moléculas inorgânicas ou íons como cálcio, potássio, etc., bem como os próprios genes, nucleotídeos, aminoácidos e mais uma infinidade de outras moléculas que são indispensáveis para a manutenção de qualquer organismo vivo.

Podemos dizer que a célula apresentada na figura 1.1 pode ser reduzida a uma rede com diversos tipos de interações, como por exemplo: gene-proteína, proteína-proteína, gene-metabólito, gene-gene, proteína-metabólito, metabólito-metabólito (figura 1.2). Este conjunto de interações recebe o nome de **interatoma** [TOY 07]. A complexidade desta rede é muito elevada e não é simples prever o que acontece com o metabolismo celular caso um gene seja deletado ou sofra mutação.

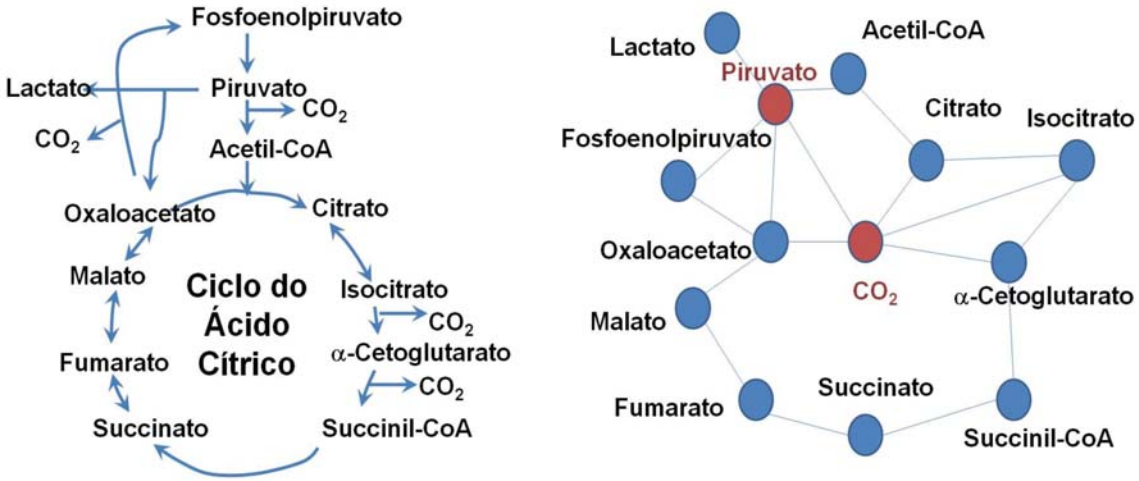


Fig. 1.2: *Ciclo do ácido citríco. (A) Representação da Rota Metabólica [NEL 04], em (B) vemos a mesma rota, porém usamos uma representação em forma de rede.*

A questão que se coloca, então, é compreender o funcionamento celular de uma maneira suficientemente profunda que possibilite a descrição do metabolismo, diagnósticos a respeito do estágio e performance celular e finalmente eventuais intervenções visando terapias ou manipulações com objetivos específicos. Assim, o objeto do estudo tem a ver com o desempenho das funções celulares. No entanto, dada a complexidade da rede de interações entre os diferentes componentes celulares, o problema está longe de ser trivial. Dois pontos de vista opostos poderiam ser adotados, em princípio. Em uma ponta, poderíamos considerar o papel exercido pelos genes ou proteínas individualmente. Na outra ponta estaria uma abordagem completamente holística em que a rede como um todo seria considerada. Como existem algumas terapias bem sucedidas que intervêm no funcionamento de alguns poucos genes, temos indícios de que podemos identificar alguns módulos funcionais atuando com relação a algumas funções biológicas específicas. Por outro lado, em muitos casos, a descrição

do funcionamento celular considerando genes isoladamente não tem apresentado resultados satisfatórios. O ideal é procurar por módulos funcionais usando o conjunto de interações, de maneira que os genes destes módulos interajam mais fortemente entre si do que com o resto dos genes dos organismos, fornecendo uma estimativa de como eles interagem com os demais módulos da rede. A performance celular poderia então ser estimada associando-se dados de transcrição de cada módulo funcional.

Neste trabalho restringimo-nos à análise de interações do tipo proteína-proteína. O Proteoma é o conjunto de proteínas de um dado organismo, mas a informação que pode ser extraída dele vai muito além da simples listagem de proteínas. Podemos intuir e deduzir informações substanciais quanto à organização e à dinâmica dos processos metabólicos, de sinalização e regulatórios de uma célula.

A análise proteômica [LOS 07, BAR 07, HID 09, GOH 07], onde mede-se o quanto cada proteína está sendo produzida por uma célula, pode mostrar como processos bioquímicos se modificam para estados patológicos e como podemos manipulá-los, mediante a administração de medicamentos, terapia gênica, etc.. Portanto, uma das grandes possibilidades da investigação dos proteomas, além do mapeamento das rotas metabólicos celulares, é a identificação de novos alvos farmacológicos, novas moléculas bioativas e marcadores biológicos que podem ser usados para diagnósticos clínicos, por exemplo, biomarcadores para câncer.

No que segue vamos primeiramente discutir abordagens matemáticas apresentadas na literatura, bem como algumas medidas experimentais, descrever alguns bancos de dados públicos e então apresentar em detalhe os objetivos da tese de doutoramento.

1.2 Redes

Podemos encontrar redes complexas nos mais diversos contextos, como por exemplo: redes sociais, biológicas, tecnológicas, informação, etc. Redes sociais podem ser formadas por grupos de pessoas que estão ligadas via relações profissionais, familiares, amizade, sexual, ou outros tipos. As redes tecnólogicas são aquelas criadas pelo o homem, como redes de esgoto, água, luz, estrada, aeroportos, etc. Redes de informação dizem respeito a mensagens como por exemplo, as redes de TV, internet, citações, e consistem de inúmeros tipos. E por

fim temos os exemplos das redes biológicas, como cerébro que é formado por uma rede de neurônios muito conectados entre si, o sistema vascular que pode ser considerado como uma rede, ou o genoma. Existem alguns modelos matemáticos de redes que podem ser usados para modelar algumas redes reais [NEW 06, PO 01, GER 95]. Uma rede é constituído por vértices e arestas. Na figura 1.2 cada metabólito foi transformado em um vértice e as reações, em arestas. Cada aresta deve ser atribuída a dois vértices; neste caso desprezamos as direções das reações.

Podemos definir a matriz de adjacência (A) como uma representação matricial de uma rede, onde cada elemento da matriz a_{ij} assume o valor 1(0) no caso de haver (ou não) uma aresta ligando os nós rotulados por i e j. Observe que esta matriz não traz informação sobre um possível direcionamento das ligações (i age sobre j mas j não age sobre i) nem sobre a intensidade dessas interações. Como exemplo, apresentamos na figura 1.3 uma rede com nós e ligações. A matriz de adjacência desta rede tem a forma da matriz 1.1.

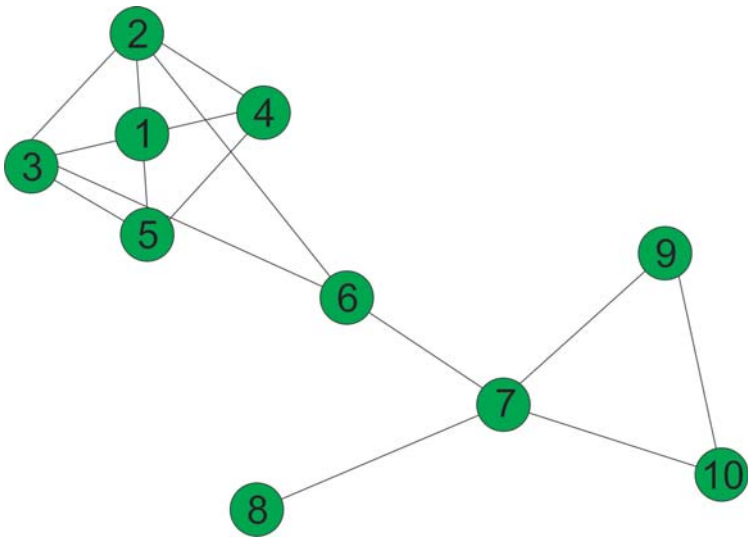


Fig. 1.3: *Rede de interação*

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{1.1}$$

1.3 Medidas

1.3.1 Conectividade e distribuição de conectividades

A conectividade k_i de um vértice i é definida como o número de arestas do vértice i. Podemos obter a conectividade a partir da matriz de adjacência:

$$k_i = \sum_{j}^{N} a_{ij}, \tag{1.2}$$

onde N é o número total de vértices, e a_{ij} é o elemento da matriz de adjacência (A). Através desta medida podemos definir também a conectividade média da rede $\langle k \rangle$:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i. \tag{1.3}$$

Com a conectividade podemos definir a distribuição de conectividades $p(k)$, que dá a fração do número de vértices da rede com conectividade k.

1.3.2 Coeficiente de clusterização

O coeficiente de clusterização[1] mede a razão entre o número existente de arestas n entre os vizinhos de um dado vértice i e o número máximo possível destas arestas:

$$C_i = \frac{2n}{k_i(k_i-1)} = \frac{1}{k_i(k_i-1)} \sum_{j=1}^{N} a_{ij} \sum_{m=1}^{N} a_{jm}a_{mi}. \tag{1.4}$$

Quando C_i é igual a zero, os vizinhos do vértice i não possuem conexão entre si, e no caso de $C_i = 1$ todos os vizinhos de i estão conectados entre si.

Assim como na conectividade, podemos ter um valor do coeficiente de clusterização médio para uma rede:

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{1.5}$$

1.3.3 *Overlap* Topológico

Esta é uma medida proposta por Barabási e Ravasz [RAV 02, RAV 03], com a qual podemos medir o grau de compartilhamento de vizinhança entre dois vértices. Trata-se de uma medida complementar à clusterização pois mesmo quando dois vértices não possuem uma aresta ligando-os diretamente, eles podem apresentar um valor de *overlap* topológico diferente de zero se tiverem algum vizinho em comum. O *overlap* topológico é definido como a fração de vizinhos compartilhados entre dois vértices i e j em relação ao número máximo possível. O *overlap* topológico é uma matriz quadrada O, onde as colunas i e linhas j são indexadas pelos vértices e cada elemento da matriz pode ter valor no intervalo $[0, 1]$, dado por:

$$O_{ij} = \frac{\Theta(a_{i,j}) + \sum_{l=1}^{N} a_{il}a_{jl}}{\min(k_i, k_j) + 1 - \Theta(a_{i,j})}. \tag{1.6}$$

Na Equação 1.6 os termos a_{il} e a_{jl} são os elementos da matriz de adjacência k_i e k_j são as respectivas conectividades dos elementos i e j. Quando i e j são vizinhos deve-se adicionar 1 ao numerador, por essa razão utilizamos $\Theta(a_{i,j})$, que é uma função degrau.

Nas referências [RAV 02, RAV 03], Barabási e colaboradores propuseram um algoritmo de hierarquização baseado na matriz de *overlap* topológico, que vamos chamar de agru-

[1] clusterização é um neologismo (do inglês *clustering*), ele tem sentido de aglomeração, agrupamento. Nesta tese será usado como "clusterização" para não perder o sentido técnico.

pamento hierárquico. O algoritmo, inspirado em algoritmos de formação de dendogramas [CAM 65] em redes filogenéticas, tem por objetivo agrupar hierarquicamente os vértices usando algum critério de semelhança ou proximidade. Barabási e colaboradores tomaram o *overlap* topológico como critério. O método do *overlap* topológico inicia por encontrar o par de vértices, digamos (u, v), com maior valor de *overlap*. A seguir constrói-se uma nova rede, agora com $(N - 1)$ vértices, em que os vértices u e v são eliminados e introduz-se um novo vértice, denotado por (u, v). O novo vértice (u, v) está conectado com o resto da rede de tal maneira que a nova matriz de *overlap* topológico, O^* pode ser obtida da antiga matriz O eliminando-se as linhas e colunas referentes aos nós u e v e introduzindo uma nova linha e uma nova coluna (u, v) com elementos dados por

$$O^*_{(u,v)w} = \frac{k_u O_{uw} + k_v O_{vw}}{k_u + k_v}, \tag{1.7}$$

onde k_u e k_v são o número de vizinhos a u e v. Esse processo é repetido até que a matriz final tenha um tamanho 1×1. No processo os vértices serão agrupados por intensidade de *overlap*, hierarquicamente. Um resultado possível, em uma rede de 10 elementos poderia ser

$$((6, (2, 3)), ((5, (1, 4)), ((9, 10), (7, 8))))$$

Neste agrupamento, os pares (2,3), (1,4), (9,10) e (7,8) foram primeiramente agrupados e então o novo elemento (2,3) foi agrupado com o vértice 6, etc. Assim, a partir desta listagem podemos reconhecer os vértices e grupos de nós com maior *overlap*, de uma forma hierárquica.

Poderíamos, mais ainda, ordenar os vértices como $6, 2, 3, 5, 1, 4, 9, 10, 7, 8$, simplesmente lendo diretamente do resultado a ordem com que os vértices aparecem. O intuito seria reconhecer vértices mais próximos nesta linha como vértices com maior *overlap* topológico. A correlação entre proximidade em um ordenamento e *overlap* é uma propriedade relevante para quantificar o desempenho de redes dinâmicas, como veremos em seções mais adiante. Em todo o caso, como seguindo este método em cada agrupamento poderíamos rotular o novo vértice por (u, v) mas também por (v, u), outros ordenamentos também seriam

compatíveis com o resultado acima. Por exemplo, $6, 3, 2, 4, 1, 5, 8, 7, 10, 9$. Assim, há uma degenerescência de $2^{(N-1)}$ nos ordenamentos possíveis, cada fator dois correspondendo a um dos parênteses do resultado final.

1.3.4 Rede Aleatória

Em 1960, Erdös-Rényi (ER) [ERD 60] criaram um modelo de rede aleatória, iniciando a partir de uma rede com N vértices e considerando uma probabilidade p que um dado par de vértices seja conectado. Este processo gera uma rede com aproximadamente $pN(N-1)/2$ arestas distribuídas aleatóriamente como na figura 1.4 **Aa**. A distribuição da conectividade segue uma lei de distribuição binomial, o que indica que muitos vértices terão a mesma quantidade de arestas, enquanto que alta e baixa conectividades são raras como se pode observar pelo gráfico na figura 1.4 **Ab**. O coeficiente de clusterização é independente da conectividade do vértice, como é visto pelo gráfico $c(k)$ versus k, onde temos uma linha horizontal ($C(k) \approx p$), figura 1.4 **Ac**.

1.3.5 Rede Livre Escala

Esta rede é obtida por meio do modelo de crescimento de redes proposto por Barabási e colaboradores [BAR 03] que, iniciando com um pequeno conjunto de vértices todos ligados uns com os outros, novos vértices são acrescentados um a um. As ligações de um novo vértice com os já existentes são criadas com uma probabilidade dada por

$$p(k_i) = \frac{k_i}{\sum_{j=1}^{N} k_j} \langle k \rangle \tag{1.8}$$

onde k_i é a conectividade do vértice e $\langle k \rangle$ a conectividade média desejada. A rede é mostrada na figura 1.4 **Ba**. Esta rede apresenta uma distribuição de conectividade $p(k)$ na forma de uma lei de potência caracterizada por um expoente $\gamma = -3$, mostrada na figura 1.4 **Bb**. A probabilidade de existir vértices altamente conectados é estatisticamente mais significativa do que na rede aleatória e tais vértices são chamados de "hubs".

A figura 1.4 **Bc**, mostra o comportamento do coeficiente de clusterização com respeito à conectividade, vemos que é uma linha reta paralela ao eixo da conectividade. Isto significa

que todos os vértices têm sempre um valor de coeficiente de clusterização muito próximo um dos outros, não importando a sua conectividade, logo não existe uma formação de módulo. No caso da 1.4 **Cc**, vemos que o coeficiente de clusterização tende a ser muito alto para vértices pouco conectados e baixo para vértices muito conectados (reta decrescente). Existe formação de módulo, vértices com baixa conectividade tendem a se ligar com seus semelhantes. Se tivéssemos uma curva gaussiana no lugar da reta decrescente na figura 1.4 **Bc**, poderíamos dizer que existe um tamanho preferencial de módulo, dado pela média da gaussiana.

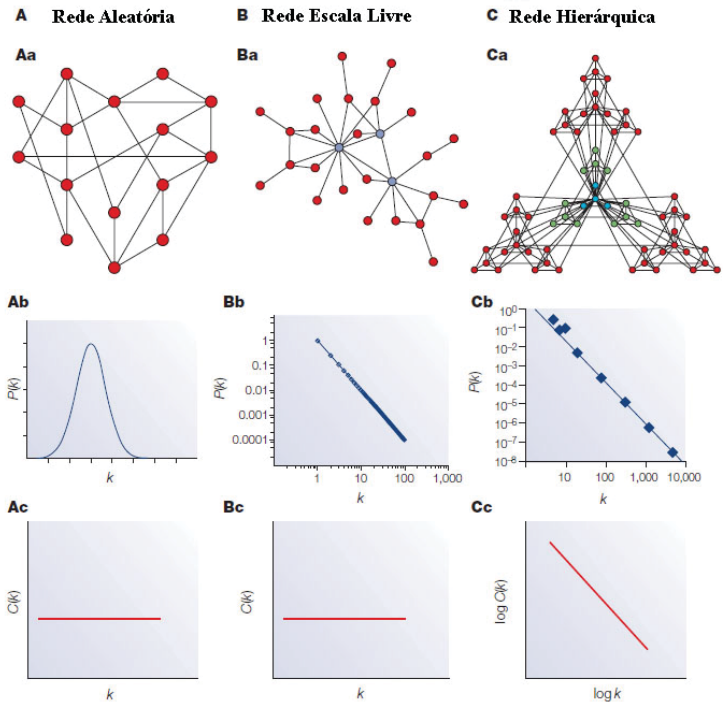


Fig. 1.4: *Principais modelos de redes, figura retirada do artigo do Barabási [OLT 04].*

1.3.6 Rede Hierárquica

Uma rede hierárquica [BAR 01] pode ser construída partindo um bloco de N vértices todos ligados entre si, que é replicado m vezes. Uma rede aumentada é gerada ligando cada vértice central destes m módulos ao vértice central do módulo inicial, formando um super módulo de $(m+1)N$ vértices. Repetindo este processo *ad infinitum* obtemos uma rede hierárquica, como na figura 1.4 **Ca**. A rede hierárquica apresenta uma topologia livre de

escala mas com uma estrutura modular. A distribuição de conectividades desta rede é uma de lei de potências com expoente $\gamma = -2,26$, como na figura 1.4 **Cb**. O interessante é que o coeficiente de clusterização escala com uma lei de potência que segue $C(k) \approx k^{\gamma}$, como pode ser visto na figura 1.4 **Cc**.

A arquitetura hierárquica apresenta vértices pouco conectados fazendo parte de áreas altamente clusterizadas com comunicação entre os vizinhos nos mais variados níveis de cluterização sendo mantidos por poucos nós muito conectados: os "hubs".

1.4 Bancos de Dados

1.4.1 STRING

STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) é um banco de dados de interação protéica, sendo mantido pelo Laboratório Europeu de Biologia Molecular (European Laboratory for Molecular Biology - EMBL) desde 2000 [vM 05, vM 07, JEN 09], disponível no sítio: *http://string.embl.de/*. Atualmente encontra-se na versão 8.3 (neste trabalho foi utilizada a versão 8.0), possui dados de 630 organismos, cerca de 2,5 milhões de proteínas catalogadas com quase 90 milhões de interações protéicas. O STRING reúne e organiza a informação existente sobre os genes, proteínas e suas interações obtida tanto de forma da pesquisa do EMBL como nos dados publicados em revistas científicas e outros bancos de dados, ver tabela 1.1. Assim reúne o resultado do trabalho de diferentes pesquisadores e laboratórios.

Mais ainda, o STRING se propõe a ser um repositório da informação que pode servir para diferentes fins, desde um estudo sociológico de quais genes seriam mais estudados em conjunto até o mais natural que é o estudo da dinâmica e metabolismo de diferentes organismos. A parte interessante deste banco de dados disponibilizado pelo EMBL é justamente o controle da qualidade e origem da informação. Assim, escolhendo um organismo, por exemplo, podemos formar uma rede de interação em que as proteínas são os vértices e as interações são listadas controladamente, escolhendo-se níveis de confiança que controlem falsos positivos e falsos negativos adequadamente.

Sigla	Nome do Banco de Dados	endereço eletrônico
BIND	Biomolecular Interaction Network Databank	http://www.bind.ca
BIOCARTA	Biocarta All Pathways	http://www.biocarta.com
BIOCYC	BioCyc collection of Pathway / Genome Databases	http://biocyc.org
DIP	Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu
GO	Gene Ontology	http://www.geneontology.org
GRID	General Repository for interaction Datasets	http://www.thebiogrid.org
HPRD	Human Protein Reference Database	http://www.hprd.org
INTACT	Intact	http://www.ebi.ac.uk/intact
KEGG	Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg
MINT	Molecular Interaction Database	http://mint.bio.uniroma2.it
PID	Pathway Interaction Database	http://pid.nci.nih.gov
REACTOME	Reactome	http://www.reactome.org
PUBMED	PubMed	http://www.ncbi.nlm.nih.gov/pubmed

Tab. 1.1: *Tabela dos bancos de dados que fazem parte do STRING.*

O STRING considera duas categorias de relações entre proteínas: interações físicas (diretas) e associativas (indiretas), que são dados provenientes de rotas metabólicas. Estas categorias são subdividas em quatro tipos de informações: Contexto Genômico, Alta-Performance, Co-Expressão e Conhecimento Prévio. Estes, por sua vez, são subdivididos em sete métodos de obtenção dos dados: *Neighborhood*, *Gene Fusion*, *Co-Ocurrence*, *Co-expression*, *Experimentos*, *Database* e *Textmining*. Tais métodos estão explicados no que segue.

- **Neighborhood, Gene Fusion e Co-ocurrence**: todos os três tipos têm o obje-

tivo de identificar pares de genes que parecem ter sofrido pressão seletiva em comum durante o processo de evolução e que hoje funcionam de maneira associada.

- **Coexpression:** aponta pares de genes que estão sendo co-expressos no mesmo organismo.

- **Experimentos:** é uma lista de interação proteína-proteína retirada de outros bancos de dados de interação protéica tais como: BIND, DIP, GRID, HPRD, IntAct, MINT, PID.

- **Databases:** é idêntico aos "Experimentos", porém são retirados de banco de dados de rotas metabólicas, estes dados são avaliados por especialistas. Exemplos: Biocarta, BioCyc, GO, KEGG, Reactome.

- **Textmining:** fornece uma lista de grupos de associação protéica que é retirada dos resumos de artigos que estão armazenados no PUBMED. Se dois genes ou proteínas estão citados no resumo, mesmo que não possuam nenhum tipo de interação, é o suficiente para dizer que existe uma associação entre elas.

O STRING nos possibilita combinar todos os métodos para a composição de uma rede de interações proteína-proteína, pela escolha de um escore, dado por

$$S = 1 - \prod_{i=1}^{7}(1 - S_i), \tag{1.9}$$

onde S é o escore ponderado sobre os sete métodos de predição, mas pode-se escolher o conjunto de métodos a serem considerados, S_i é o escore individual de cada um dos métodos escolhidos.

Este cálculo é feito para cada par de interação. As redes que serão utilizadas neste trabalho foram obtidas usando somente seis métodos, foi excluído o *Textmining*, e o escore utilizado foi de 0.8, para equilibrar os números de pares falsos positivos e falsos negativos. As probabilidades (scores) invididuais são calculados tomando como base o KEGG [KAN 02]. A razão disso é que no KEGG os dados são analisados (curados) manualmente, vários organismos são disponibilizados e diferentes áreas funcionais são consideradas. Mais ainda, o KEGG está em constante atualização sendo portanto, indicado como *padrão ouro* para

verificação da existência ou não de uma associação protéica. Os escores fornecem informação sobre a probabilidade de, dado que o STRING encontre associação entre duas proteínas, elas interajam em uma mesma rota metabólica do KEGG.

1.4.2 *Gene Ontology* (GO)

A ferramenta *Gene Ontology* (GO) [ASH 00], está disponível no sítio (*http://www.geneontology.org/*) e tem como objetivo classificar e organizar as descrições dos genes e seus respectivos produtos. Considera várias espécies e usa informação de outros bancos de dados para isto. O GO iniciou em 1998 com somente três organismos: *Saccharomyces cerevisiae*, *Mus musculus* e *Drosophila melanogaster*. Atualmente tem mais de 50 espécies catalogadas.

As informações foram organizadas em três classificações indepentes - as ontologias - que seguem critérios diferentes. Cada gene e seu respectivo produto está anotado nas três ontologias que são independentes de organismo. As ontologias estão organizadas hierarquicamente, com classes, subclasses, sub-subclasses, etc.

Ontologias

- **Componente Celular:** o critério de classificação é a localização, ou seja, onde os produtos dos genes atuam. Esta ontologia não se refere a atividades, e sim a local de atuação. Exemplos: célula (básica unidade estrutural e funcional de todos os organismos), organelas (estrutura organizada e de morfologia e função distinta: ribossomos, vacúolos, mitocôndria, núcleo, etc.).

- **Função Molecular:** as funções de um gene são as atividades que ele pode realizar tais como transportar substâncias, ligar-se a produtos, fixar-se a produtos, ou modificar moléculas. Podemos fazer uma analogia com uma montadora de automóveis, onde os indivíduos têm diferentes habilidades (funções), mas trabalham juntos com objetivos diferentes.

- **Processo Biológico:** tal processo tem um início e um fim bem definidos, como por exemplo, divisão celular, transcrição, o ciclo de Krebs, etc. O exemplo anterior da fábrica automotiva como um todo pode ser encarado como um processo biológico. Cada

processo biológico pode haver funções de transporte, fixação, etc (que corresponderiam a diferentes funções moleculares) e para cada processo as etapas podem ser realizadas em diferentes localizações.

Cada Ontologia tem a estrutura de um grafo direcionado não-cíclico, cada termo possui um ou mais termos ancestrais e termos filhos. Cada termo pode ter um ou mais produtos anotados.

Esta ferramenta pode ser utilizada de diferentes maneiras, mas a essência é que lista os genes ou produtos gênicos que atuam em um dado compartimento celular, que executam uma determinada função molecular ou que são responsáveis por um processo biológico.

1.4.3 KEGG

Enciclopédia de Kyoto dos Genes e dos Genomas (KEGG) [KAN 02, OKU 08] está disponível no sítio (*http://www.genome.jp/kegg/*) e foi criada em 1995 pelo programa japonês do genoma humano. O objetivo é automatizar os passos para interpretação do significado biológico codificado nos dados obtidos pela HUGO. A KEGG visa informatizar os conhecimentos da genética, bioquímica e biologia molecular e celular em termos de interações moleculares e/ou gênicas. O ponto forte deste banco de dados é o KEGG PATHWAYS, que é uma coleção de imagens gráficas para as rotas bioquímicas. Existem cerca de 340 mapas de referência para as rotas metabólicas que são elaboradas manualmente e continuamente atualizadas de acordo com as evidências coletadas, como por exemplo a figura 1.5 mostra um mapa da rota bioquímica para o ciclo do citrato.

KEGG tem sido amplamente utilizado como uma base de conhecimento biológico de referência para a interpretação de grandes conjuntos de dados como, por exemplo, dados gerados por microarranjo (*microarrays*). Atualmente é formado por 16 bancos de dados, que estão divididos em três categorias: Sistemas de Informação, Informação Genômica e Informação Química. Informação Genômica e Química representam os blocos de construção molecular da vida, enquanto que Sistemas de Informação são os aspectos funcionais do sistema biológico, como a célula e o organismo que são construídos a partir dos blocos moleculares. KEGG é uma representação computadorizada dos sistemas biológicos. Ele é

baseado no conceito de grafos direcionados para representação e manipulação de objetos de diferentes níveis moleculares.

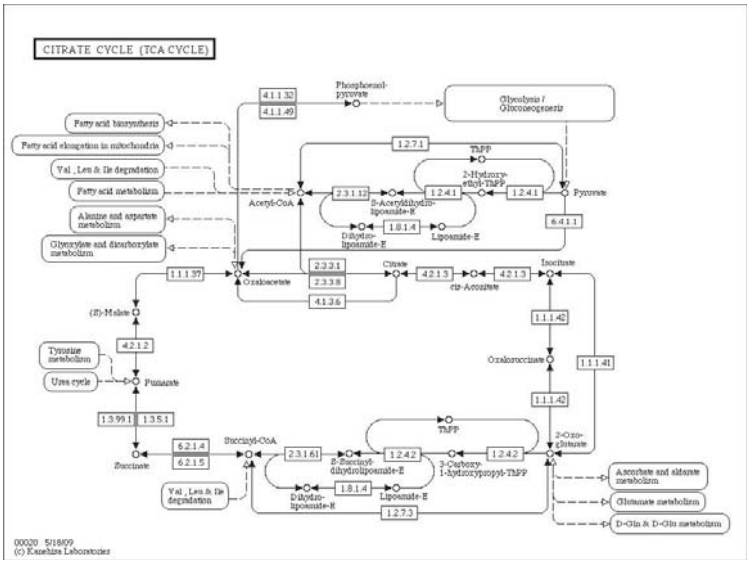


Fig. 1.5: *mapa de uma rota metabólica obtida do KEGG.*

1.4.4 David Tools

O Banco de dados DAVID *(Database for Annotation, Visualization and Integrated Discovery)* [HUA 07, HUA 09] está disponível no sítio (*http://david.abcc.ncifcrf.gov/*) e consiste de um grupo de ferramentas para análise de grandes quantidades de genes visando auxiliar no entendimento do significado biológico de tais elementos. O usuário pode entrar com uma lista de gene, selecionar "Anotação Funcional", e identificar os termos do GO aos quais os genes pertecem, assim como rotas metabólicas do BioCarta e KEGG, genes homólogos, associações gene-doenças, artigos, características de sequências gênicas, domínios funcionais de proteínas, etc.

Em especial, a ferramenta "Classificação Funcional" fornece um método ágil para organizar grandes listas de genes em grupos funcionalmente relacionados para ajudar a desvendar o conteúdo biológico obtido pelas tecnologias de alto rendimento (microarranjo).

1.4.5 Gene Expression Omnibus (GEO)

.

Novas tecnologias possibilitam aos cientistas analisar a expressão de inúmeros genes em uma única experiência com alta rapidez e eficiência.

O microarranjo de DNA ou DNA-chip, consiste num arranjo pré-definido de moléculas de DNA que podem ser fragmentos de DNA genômico, DNA complementar, etc. Tais fragmentos são quimicamente ligados a uma superfície sólida, geralmente lâminas de microscópio cobertas com compostos que adquirem carga positiva. Os microarranjos também podem ser preparados em membranas de nylon positivamente carregadas [GU 06].

Os microarranjos são utilizados na detecção e quantificação dos níveis de expressão de ácidos nucléicos provenientes de amostras biológicas, os quais são postas para hibridizar com o DNA fixado no arranjo. Estas amostras são o objeto de estudo do pesquisador, sejam elas preparadas a partir de células normais, cancerosas, tecido neuronal, etc. É extraído o RNA mensageiro dos tecidos, que depois é convertido em DNA complementar.

O DNA complementar é marcado com fluorocromos de cianina 3 (Cy3) ou cianina 5 (Cy5) quando se utiliza microarranjos vidro, ou é usado o isótopo de fósforo-33 quando os microarranjos são preparados em membranas de nylon. Estes tipos de marcação tornam a detecção viável.

Para ambos os casos é necessário a geração de uma imagem de hibridização, que é obtida por meio de leitores. No caso de fluorocromos é usado um leitor a laser e leitores de fósforo para arranjos tratados com o fósforo 33. Os leitores medem a quantidade de fluorocromos em cada sonda do microarranjo, que possui um fragmento de DNA que irá hibridizar com os fragmentos específicos de DNA contido na amostra.

O uso mais frequente dos microarranjos é na determinação da expressão gênica (perfil do transcriptoma. O diferencial desta técnica é tornar possível determinar a expressão de milhares de genes num único experimento, em intensidades que correspondem à quantidade relativa dos produtos expressados pelas células. Para a preparação dos microarranjos são utilizados robôs altamente precisos que aplicam as diferentes amostras de DNA em diminutos pontos (spots) no centro de uma lâmina de microscópio com a superfície quimicamente

preparada, cuja densidade aproximada de 10.000 pontos/cm^2. Cada ponto representa um segmento gênico em particular. Quanto mais pontos (sondas) no microarranjo, mais abrangente ele será na análise do trancriptoma.

Nos últimos anos a tecnologia de microarranjo teve uma explosão de aplicações que vão muito além de analisar os níveis de expressão do gene. Exemplos: a análise do polimorfismo do nucleotídeo único (SNP) ou o nível de metilação do genoma. Muito desta informação está espalhada em toda a Internet ou ainda não está disponível para o público. Esta avalanche de dados exige padronização de armazenamento, compartilhamento e técnicas de edição. Para apoiar o uso público e divulgação de dados de expressão gênica, o Centro Nacional para Informação Biotecnológica (NCBI) dos EUA lançou o *Gene Expression Omnibus* (GEO).

O GEO [EDG 02] (*http://www.ncbi.nlm.nih.gov/geo/*) é o maior repositório público de dados de expressão gênica obtidos por experimentos de alto rendimento (*high throughput*). Estes dados são gerados pela comunidade científica usando tecnologias de alto rendimento, principalmente microarranjos. Além de arquivamento de dados, GEO fornece ferramentas para auxiliar os usuários de todos os tipos a pesquisar rapidamente, analisar, visualizar e baixar esses dados. Hoje, o banco detém mais de 15.000 experimentos que compreende 450.000 amostras depositadas por laboratórios de todo o mundo.

1.5 Objetivos

Este trabalho tem como objetivo propor uma metodologia de ordenamento de genomas em uma lista a partir das interações protéicas. Este ordenamento possibilita definir módulos funcionais, e caracterizá-los sabendo quais processos biológicos estão associados a eles, e projetar dados de expressão gênica.

As redes são extraídas do STRING e ordenadas, as medidas de redes tais como a conectividade e a clusterização podem ser úteis na definição de módulos, impondo que módulos devem possuir uma alta conectividade e coeficiente de clusterização.

O *David tools* fornece a informação do enriquecimento funcional de cada módulo no ordenamento. Com o uso do *KEGG* e do *Gene Ontology* é possível extrair os dados de processos biológicos e rotas metabólicas, e assim plotá-los sobre o ordenamento. Este passo

é fundamental para entender como os processos biológicos estão agrupados, em que pontos do ordenamento eles estão situados, e como eles estão interagindo uns com os outros.

A projeção dos dados de expressão gênica sobre o ordenamento, revelando a atividade transcricional dos módulos funcionais já identificados pode mostrar diferenças no metabolismo de células sadias e células cancerosas, a ação de um determinado farmáco sobre um conjunto de tecidos, etc. O espectro obtido com os dados de microarranjo (transcriptoma) será chamado de transcriptograma (*transcriptoma + pospositivo -grama*) que significa: sinal de transcriptoma. O *transcriptograma* tem como objetivo auxiliar no entendimento do significado biológico obtido nos microarranjos.

Todos os programas desenvolvidos para análise devem ser concatenados para formar um software livre para ambientes Windows e Linux. Este software tem como objetivo auxiliar um bioquímico a executar funções para as quais seriam necessárias a ajuda de um físico e vice-versa, até mesmo um usuário sem grande conhecimento em ambas as áreas (física e bioquímica) poderia usá-lo.

Capítulo 2

Organização de Redes

Técnicas experimentais e ferramentas computacionais possibilitam a obtenção de uma quantidade crescente de dados biológicos que são depositados em bancos de dados usualmente disponibilzados publicamente. Tais bancos listam, para cada espécie, os genes, as proteínas, os metabólitos e suas respectivas interações. O atual desafio é como transformar todos estes dados em informação que leve à melhor compreensão do funcionamento de células e organismos e dêem lugar a ferramentas de diagnóstico e terapia de doenças originadas pelo mal funcionamento do metabolismo celular. Existem alguns trabalhos [RAV 02, STR 03, YOO 04, OLT 02], baseados em redes filogenéticas que tentam relacionar topologias de redes protéicas com a organização funcional. A importância de uma ferramenta para ordernar listas de genes é prover um critério para a definição de subconjuntos de genes cujo funcionamento deve ser investigado conjuntamente, e estimar os erros de separá-los do resto do genoma. Usando o método de Monte Carlo, propomos uma técnica para o ordenamento destes dados em uma lista, utilizando minimização da função custo que favorece o agrupamento de genes na lista, comparamos os resultados com o método de clusterização do *overlap* topológico proposto por Barabási e colaboradores [RAV 02].

2.1 Hierarquizando Aglomerados

2.1.1 Método de Minimização da Função Custo (CFM)

Nesta seção é sugerido um método de organização de genes em uma lista ordenada a partir da matriz de adjacência. Consideramos redes não-direcionadas e binárias.

A partir de uma rede G com N vértices e E arestas, rotulamos arbitrariamente os vértices com números inteiros no intervalo $[1, N]$ e montamos a matriz de adjacência A com N colunas e N linhas, de tal maneira que o elemento a_{ij} é igual a 1 para o caso em que os vértices i e j interagem entre si e 0 em caso contrário.

A enumeração dos vértices é arbitrária e poderia ser feita diferentemente. De fato, existem $N!$ diferentes ordenamentos possíveis da lista com N vértices. A cada um destes ordenamentos, corresponde uma matriz de adjacência. A partir de uma dada configuração da lista de vértices, podemos obter uma outra configuração pela escolha aleatória de dois genes da lista cujos lugares na lista são trocados um com o outro. A nova matriz de adjacência, que corresponde ao novo ordenamento da lista, é obtida pela troca adequada das linhas e colunas que correspodem aos vértices trocados, como é apresentado na figura 2.1. Neste exemplo temos uma matriz 10×10, correspondendo a rede da figura 1.3. A quantidade de configurações possíveis é de $10! = 3628800$. As duas matrizes da figura 2.1 representam a mesma rede, só que os vértices foram enumerados de forma diferente. Entre todas as 3628800 configurações possíveis, qual delas representa melhor os vértices em um alinhamento unidimensional? A resposta depende do objetivo de um tal ordenamento.

Podemos arranjar os vértices de tal maneira que a matriz A tenha mais elementos não nulos possíveis na sua diagonal, tornando fácil a visualização de grupos interagentes a partir da diagonal da matriz e a estimativa da interação destes grupos com o restante dos vértices. Neste caso a matriz B é uma melhor representação da rede, pois ela é mais diagonal que a matriz A.

O objetivo final do ordenamento da lista é colocar próximos os vértices que interagem, de uma maneira que possibilite o reconhecimento de agrupamentos de vértices que interagem mais fortemente entre si do que com o resto dos vértices. Mas também queremos que dois agrupamentos que interagem mais fortemente entre si estejam também mais próximos. Estas características podem ser obtidas em algum grau exigindo que as configurações da matriz de adjacência satisfaçam ao máximo duas condições que, infelizmente, nem sempre são concordantes (isto é, algum compromisso se faz necessário).

A primeira condição impõe que tanto melhor é a configuração quanto menos interfaces branco/preto a matriz de adjacência apresenta. Esta imposição favorece o agrupamento

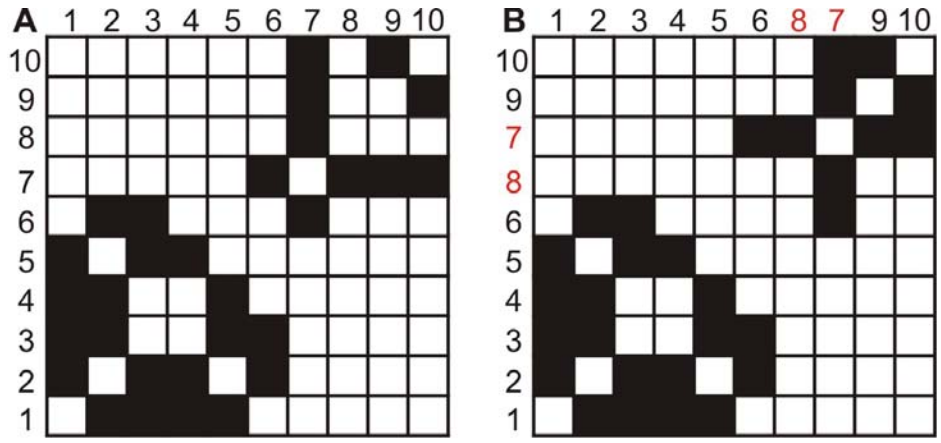


Fig. 2.1: *(A)Matriz de adjacência para uma rede de 10 vértices e 15 interações, como apresentada na seção 1.3. (B) A matriz de adjacência para a mesma rede, mas agora trocando o vértice 7 pelo 8. Neste caso, as colunas 7 e 8 foram trocadas de posição bem como as respectivas linhas.*

de vértices que tenham vizinhanças semelhantes. A segunda condição impõe que tanto melhor é a configuração quanto mais próximas da diagonal estiverem as interfaces restantes. Estas condições podem não ser concordantes quando o reposicionamento de um vértice traz algumas interfaces mais para perto da diagonal (o que é bom) ao mesmo tempo que cria interfaces adicionais (o que é ruim). Considere a figura 2.2, que apresenta um trecho de uma

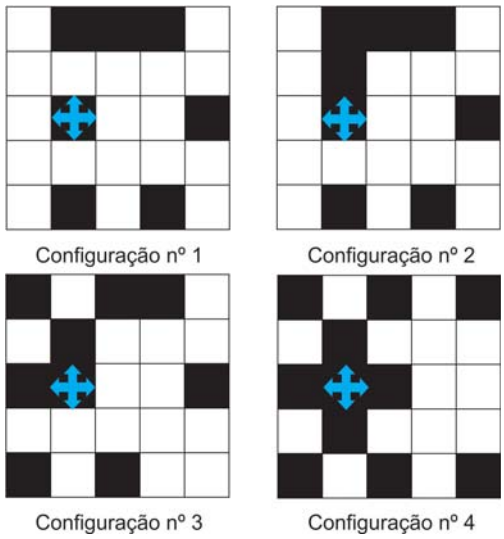


Fig. 2.2: *A figura apresenta quatro tipo de configurações possíveis.*

matriz de adjacência. Na figura apresentamos 4 situações possíveis para um píxel negro da matriz, localizado no ponto (i, j) e representando a interação entre os vértices localizados nas posições i e j da lista ordenada de vértices. Estamos considerando apenas píxels que são primeiros vizinhos. Atribuímos um custo E associada a cada elemento da matriz, de tal maneira que seja igual a soma de primeiros vizinhos diferentes. Nos exemplos das figura 2.2 temos $E_1 = 4$, $E_2 = 3$, $E_3 = 2$ e $E_4 = 0$. Usamos condições de contorno periódicas.

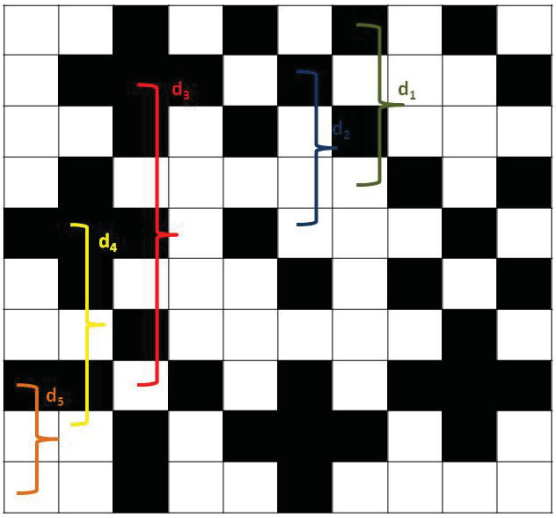


Fig. 2.3: *Representação matricial de uma rede onde é mostrada a distância do elemento não-nulo até a diagonal.*

Para assegurar a proximidade das interfaces da diagonal, multiplicamos a distância do píxel central até a diagonal pela custo de interface com os primeiros vizinhos do pixel central. Esta distância, para um píxel localizado no ponto i, j da matriz de adjacência é proporcional a $|i-j|$, como mostrado na figura 2.3. Somando esses custos sobre todos os sítios da matriz, obtemos a equação 5.1, que dá o custo de uma configuração da matriz adjacência.

Agora será utilizada uma estratégia baseada no Método de Monte Carlo para rearranjar a matriz minimizando $\mathcal{E}$.

$$\begin{aligned}\mathcal{E} &= \sum_{i=1}^{N}\sum_{j=1}^{N}|i-j|[|a_{i,j}-a_{i+1,j}|+|a_{i,j}-a_{i-1,j}| \\ &\quad +|a_{i,j}-a_{i,j+1}|+|a_{i,j}-a_{i,j-1}|]. \end{aligned} \tag{2.1}$$

O processo é iniciado com a enumeração aleatória dos vértices seguida pela obtenção da matriz A (lembrando que a randomização não irá criar e nem destruir interações) que tem um valor da função custo $\mathcal{E}_i$. A seguir são selecionados randomicamente dois vértices, os quais são trocados de posição. Obtém-se uma nova matriz e, a partir dela, é obtido um valor de custo $\mathcal{E}_f$. Se $\mathcal{E}_f < \mathcal{E}_i$ o novo estado é aceito e o processo é repetido. Caso $\mathcal{E}_f > \mathcal{E}_i$ o novo estado é aceito com probabilidade $\exp\left[-(\mathcal{E}_f - \mathcal{E}_i)/T\right]$, onde T é um parâmetro da simulação que é similar à temperatura nesta implementação de Método de Monte Carlo. Caso a troca não seja aceita, retorna-se para o estado inicial e um novo par de vértices é escolhido. A simulação pára quando é obtido um estado estável frente à qualquer troca de posições dos vértices.

Fig. 2.4: *Gráfico da Função custo versus estados de uma matriz hipotética.*

Inicialmente consideramos um valor elevado T_0 e vamos diminuindo a temperatura gradativamente da seguinte maneira. Para cada valor de T, deixamos o sistema evoluir por um determinado número de Passos de Monte Carlo (PMC), suficiente para que o sistema possa atingir o equilíbrio termodinâmico para a dada temperatura. A temperatura é então reajustada por meio de uma razão de esfriamente μ, tal que $T^* = \mu T$, sendo $0 < \mu < 1$, onde T^* é a nova temperatura. Este processo é repetido até a temperatura do sistema ser muito baixa ($T \approx 0$). Esta técnica, conhecida como *simulated annealing*, permite escapar de mínimos locais.

A evolução da organização da rede da *Saccharomyces cerevisiae* é apresentada na figura 2.5. Para o estado inicial a matriz de associação da rede é aleatorizada, esse processo não

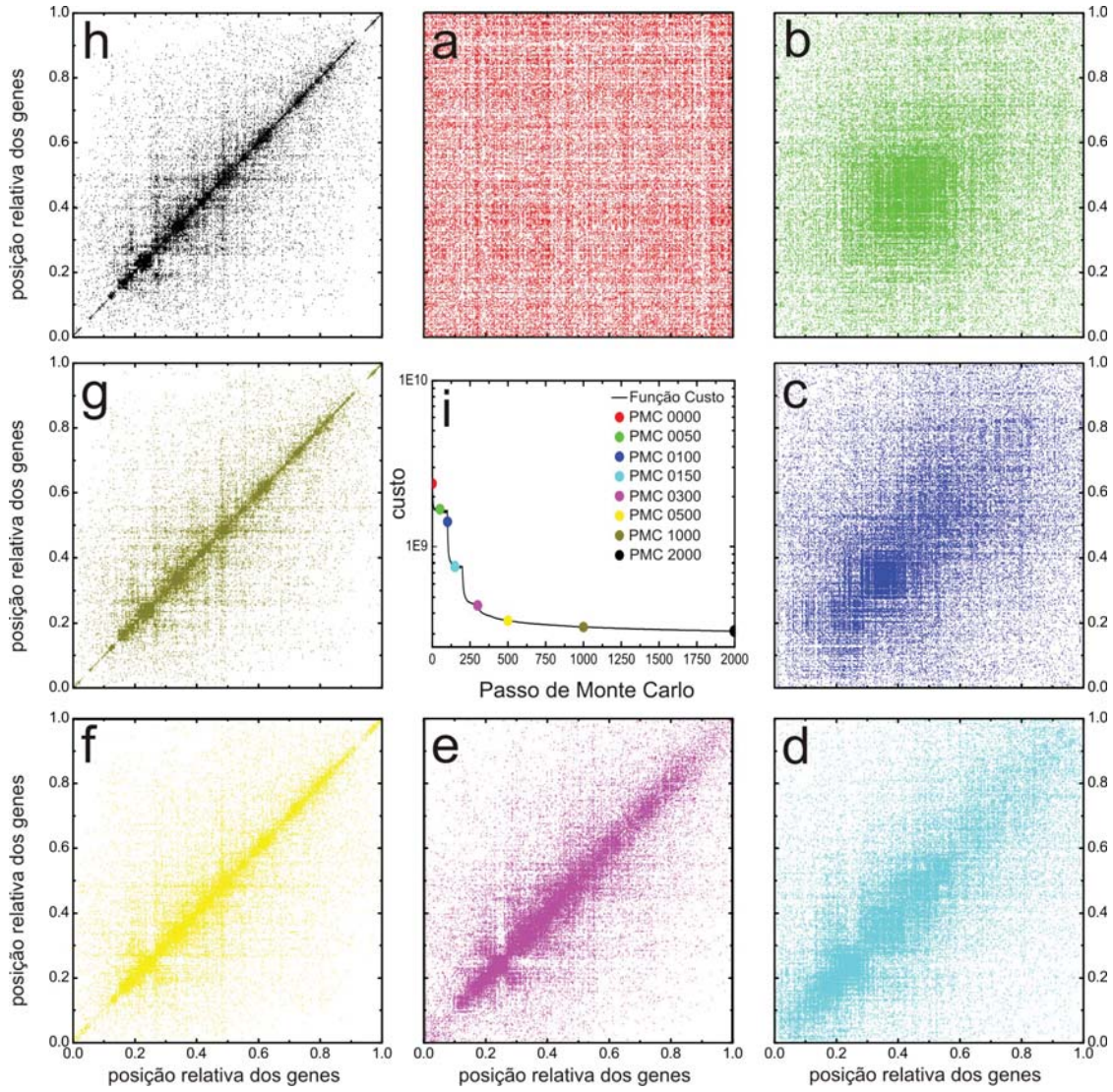


Fig. 2.5: *Painel com a evolução temporal do ordenamento da rede. (a) estágio 0, (b) estágio 50, (c) estágio 100, (d) estágio 150, (e) estágio 300, (f) estágio 500, (g) estágio 1000, (h) estágio 2000. O item (i) é um gráfico do custo versus a evolução temporal do ordenamento.*

destrói e nem gera interações. A cada passo de Monte Carlo, cada vértices i é correlacionado com os seus respectivos vizinhos (figura 2.5 **a-h**). Quanto maior a correlação menor será o custo (figura 2.5)

Este reordenamento foi aplicado em seis redes protéicas, uma para cada organismo como pode ser visto na tabela abaixo:

Organismo	número de genes	número de interações
Arabidopsis thaliana	4245	77357
Drosophila melanogaster	3746	38930
Escherichia coli	3156	12601
Homo sapiens	9019	111602
Mus musculus	5516	86469
Saccharomyces cerevisiae	4655	47915

Tab. 2.1: *tabela de espécies estudadas.*

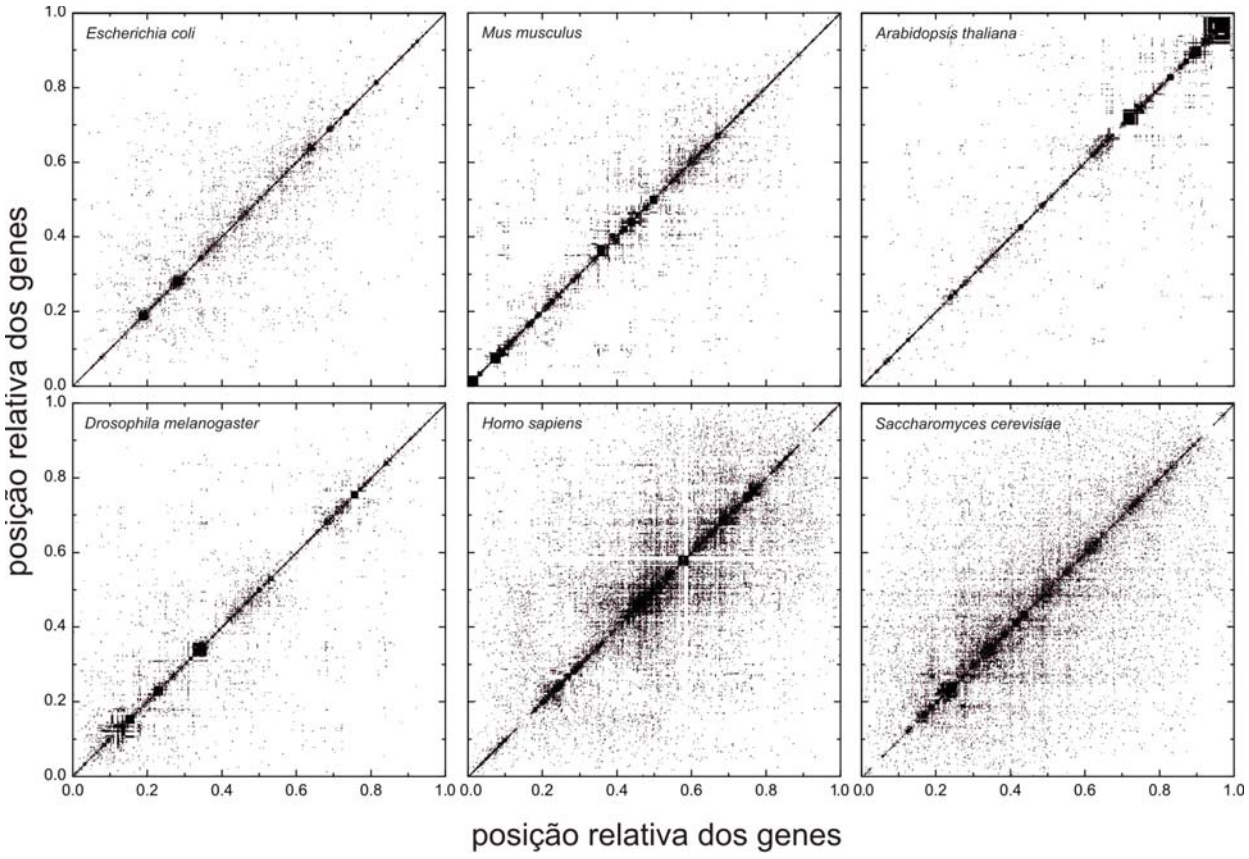


Fig. 2.6: *Este painel apresenta seis proteomas que foram reorganizados pelo CFM. Os eixos das figuras foram normalizados para uma melhor comparação.*

Tais redes foram ordenadas com o CFM, o resultado é apresentado na figura 2.6. Os eixos das matrizes foram normalizados, para uma melhor comparação. Fica evidente a formação de módulos na diagonal. As extremidades das matrizes tendem a ter poucos elementos

não-nulos enquanto que no centro da matriz existem muitos elementos: do ponto de vista da função custo é mais favorável ter vértices com conectividade baixa nas extremidades e ter os vértices mais conectados pelo centro da matriz. O fato de não impormos condições de contorno periódicas para o cálculo da distância do elemento não-nulo até a diagonal da matriz é o causador do "formato de folha" na matriz de adjacência.

Fica uma pergunta: "Como podemos extrair informações biológicas de módulos funcionais encontrados nestas matrizes de adjacência?", que será abordada nas próximas seções.

2.2 Redes Artificiais

Para ilustrar as vantagens do ordenamento e os conceitos de modularidade de janela, rugosidade e probabilidade de interação criamos quatro redes artificiais com 4655 vértices. As redes já representam um ordenamento adequado para evidenciar seu caráter modular e suas matrizes de interação são apresentadas na figura 2.7, e ali são rotuladas de **(a)** exponencial, **(b)** exponencial modular,**(c)** rede modular e **(d)** rede aleatória.

- Rede exponencial: a matriz de adjacência é construída a partir da probabilidade p, dada por:

$$p = 0.05 \exp(-\frac{|j-i|}{2150}) \tag{2.2}$$

de que um dado elemento da matriz A seja igual a um, isto é, se $a_{ij} = 1$ então as proteínas localizadas nas posições i e j da lista de proteínas interagem. Para garantir a simetria de A, sempre que $a_{ji} = a_{ij} = 1$. A rede possui 95362 interações.

- Rede exponencial modular: o mesmo que para a rede exponencial, mas com a probabilidade de interação dada por:

$$\begin{aligned} p = {} & 0.0125 \exp(-\frac{|j-i|}{\sqrt{\frac{16(i+j)(1-i-j)}{4655^2}}}) \sin^2(\frac{7\pi(i+j)\sqrt{2}}{9310}) \\ & +0.0063 \exp(-\frac{|j-i|}{\sqrt{\frac{32(i+j)(1-i-j)}{4655^2}}}) \cos^2(\frac{3\pi(i+j)\sqrt{2}}{9310}) \\ & +0.065 \exp(-\frac{|j-i|}{\sqrt{\frac{8(i+j)(1-i-j)}{4655^2}}}) \cos^2(\frac{4\pi(i+j)\sqrt{2}}{9310}) \end{aligned} \tag{2.3}$$

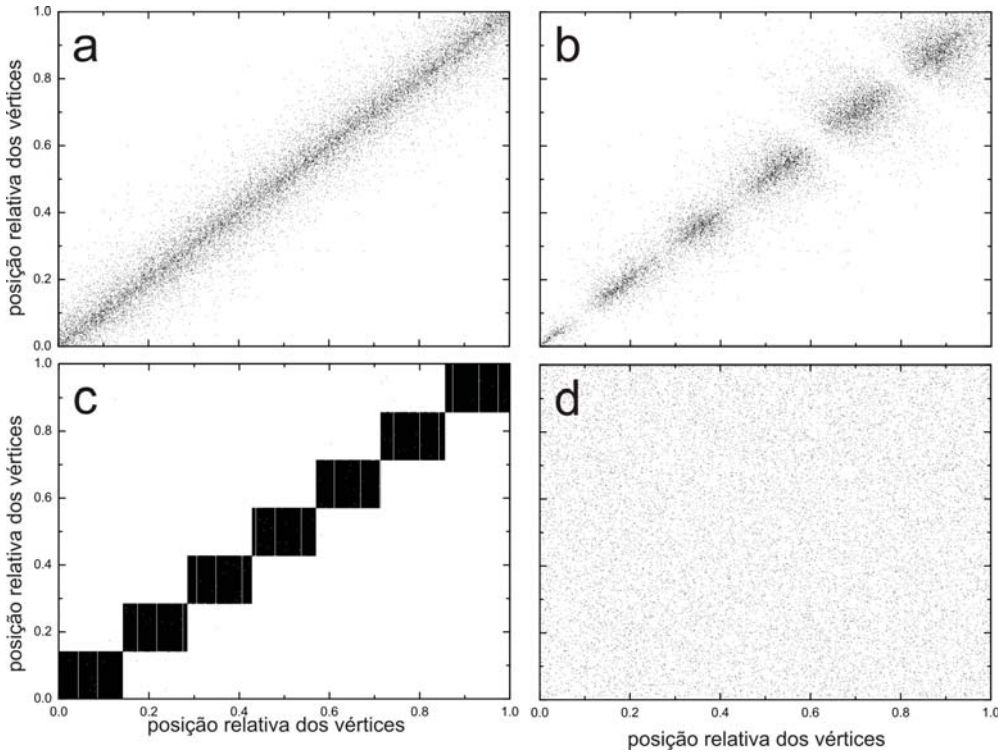


Fig. 2.7: *Painel das matrizes de adjacência para as redes artificiais. (a) rede exponencial, (b) rede exponencial modular, (c) rede modular e (d) rede aleatória.*

Observe que esta rede apresenta módulos (figura 2.7 **b**). A rede tem um total de 76352 interações.

- Rede modular: Esta rede é composta de sete módulos nos quais a probabilidade de que dois vértices interajam é de 80%. A probabilidade de que dois vértices de módulos diferentes é muito baixa, da ordem de 0.6%. A rede tem ao todo 1855862 interações.

- Rede aleatória: para a construção desta rede as interações foram escolhidas aleatoriamente, respeitando a simetria das interações e garantindo que o número de interações fosse de 49830 interações.

2.3 Modularidade de Janela

Módulo funcional [FRA 05, VIN 08] é um conjuto de genes que interagem entre si e realizam alguma função específica no metabolismo de um organismo, como por exemplo o ciclo de

Krebs. Aqui ainda não consideramos as classificações de natureza biológica dos vértices (ontologias, rotas metabólicas); encontramos os módulos por meio do rearranjo da matriz de interações, que traz informação sobre a existência ou não da interação entre dois vértices sem detalhar em quais processos biológicos tais genes participam. Para avançarmos nesta direção, foi necessário utilizar uma medida chamada de Modularidade de Janela [RF 10].

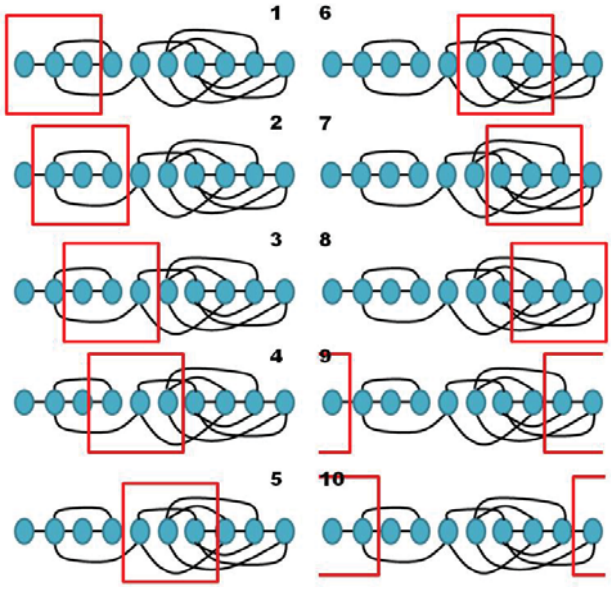


Fig. 2.8: *Cálculo da Janela de modularidade.*

Observe a figura 2.8, que representa um conjunto de genes em uma lista ordenada e onde estão indicadas as interações. Primeiramente definimos o tamanho da janela, que no caso da figura, possui tamanho 3. A janela é então um intervalo na lista ordenada. Considere uma janela centrada no i-ésimo vértice da lista. A modularidade de janela $M(i)$ do i-ésimo vértice é definida como a razão do número de arestas que unem quaisquer dois vértices contidos na janela pelo número de conexões que envolvem ao menos um vértice da janela. Passamos esta janela por todo o ordenamento usando condições de contorno periódicas, obtemos assim um perfil de modularidade. Observe que tal perfil é uma característica do ordenamento. O mesmo conjunto de vértices e interações dão lugar a perfis de modularidade diferentes se os genes estão ordenados diferentemente. Módulos muito interativos apresentam alta modularidade quando a janela é aproximadamente igual ao tamanho do módulo. Já a separação entre os módulos interativos é detectada por baixos valores de modularidade

quando a janela é pequena.

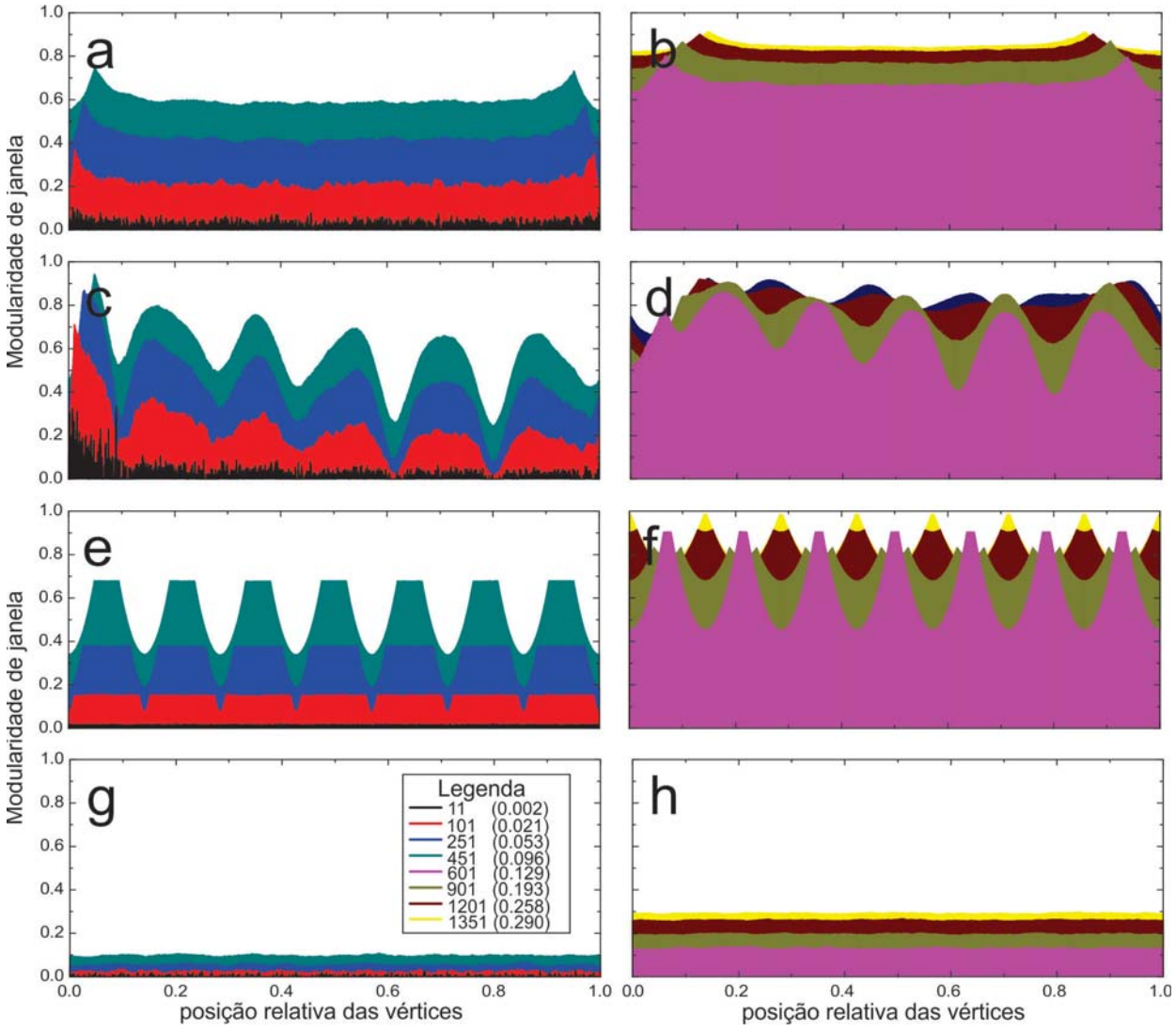


Fig. 2.9: *O painel apresenta perfis de modularidade para redes artificiais. Rede exponencial (a) e (b), rede exponencial modular (c) e (d), rede modular (e) e (f), e rede aleatória (g) e (h)*

A escolha da largura da janela é arbitrária. No caso da largura da janela ser de tamanho 1, nunca irá existir uma interação entre dois vértices dentro da janela e, o perfil de modularidade será igual a 0 ao longo do ordenamento. Se aumentamos a janela em algumas unidades, a modularidade aumenta mas ainda é baixa. Por outro lado, quando o comprimento da janela é igual ao número total de vértices da rede o perfil modularidade será igual a 1, pois todas as interações da rede estão dentro da janela.

Se aplicarmos o conceito de modularidade de janela nas redes artificiais que construímos, é possível tirar informação sobre a existência de módulos. A figura 2.9 tem 8 perfís de modularidade para cada rede. As figuras **(a)** e **(b)** referem-se à rede exponencial. A modularidade é quase constante e não há módulos. Nos gráficos **(c)** e **(d)** vemos o perfil para a rede exponencial modular. Para janela pequena a modularidade é baixa. Conforme a janela vai aumentando o nível da modularidade vai subindo e os módulos vão surgindo. Para a janela $w = 601$ os dois primeiros módulos começam a se fundir num único agrupamento, assim surge os módulos formado por módulos. A rede modular apresenta 7 módulos e, ao aplicarmos janela de largura inferior ao tamanho do módulo, o perfil apresenta uma platô no centro dos módulos com o máximo valor de modularidade. Quando a janela possuí largura maior do que o tamanho do módulo, o máximo do perfil é deslocado, como vemos para os casos de $w = 1201$ e $w = 1351$. No caso da janela $w = 1351$, o máximo ocorre quando o centro da janela está na interface dos dois módulos, pois esta janela abrange todos nós dos dois módulos. Para a rede aleatória não temos módulos, o perfil é constante.

Fica, assim, um problema: qual a melhor janela? Como definir o melhor tamanho de janela para a rede analisada? Uma possibilidade é estudar a rugosidade do perfil de modularidade para várias janelas, como veremos a seguir.

2.4 Rugosidade

A rugosidade $\mathcal{R}(w)$ é definida como o desvio padrão da modularidade calculada para janela de largura w. Se varrermos a rede para várias janelas, podemos construir uma curva do desvio padrão versus tamanho w de janela, dado por:

$$\mathcal{R}(w) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(M_w(i) - \langle M_w \rangle)^2}. \tag{2.4}$$

Quando calculamos as curvas de rugosidade para as redes artificiais, figura 2.10, vemos que a rede alealória tem uma baixa rugosidade, já que distribuição das interações na matriz de interação é homogênea e não temos formação de clusters. O mesmo acontece para a rede exponencial. Para a curva da rede exponencial modular, temos um pico de rugosidade e depois conforme a janela vai aumentando, o valor da rugosidade vai decaindo mas ainda

existe surgimento de pequenos picos que poderiam ser os módulos formados por outros módulos menores. A rede modular apresenta vários picos de rugosidade, o máximo ocorre para janela $w = 591$ e o pico secundário em torno de $w = 1301$. O valor do primeiro pico é próximo do tamanho do módulo.

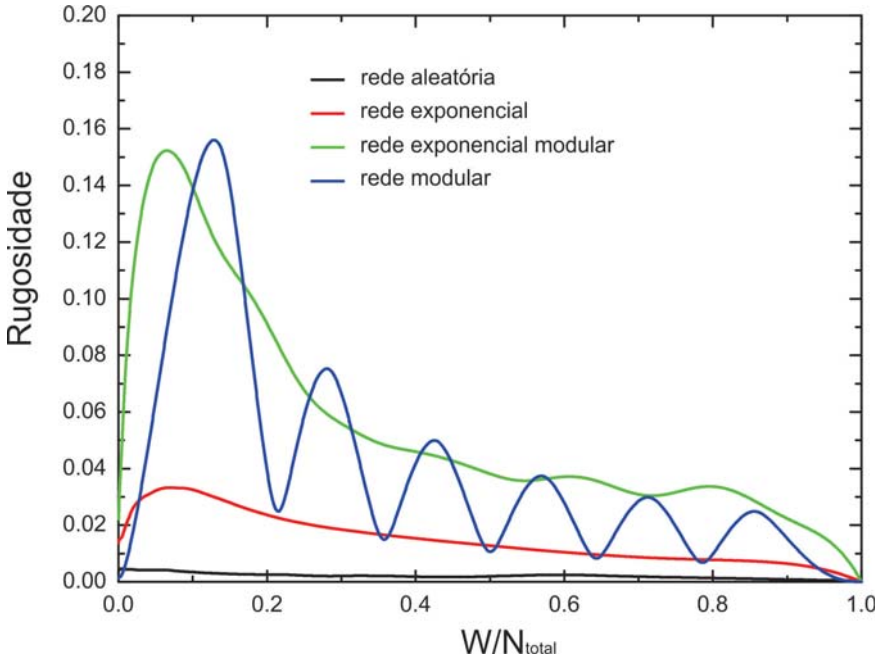


Fig. 2.10: *Curva de rugosidade para as redes artificiais.*

A figura 2.11 apresenta cinco curvas de rugosidade, uma para cada organismo ordenado. As espécies *Arabidopsis thaliana* e *Saccharomyces cerevisiae* possuem o ponto de máximo na mesma região, enquanto que *Mus musculus*, *Escherichia coli* e *Drosophila melanogaster* possuem máximo em pontos diferentes. Talvez essas similaridades sejam frutos de alguma característica independente da conectividade e do coeficiente de clusterização, isso requer mais estudo.

Mostramos no gráfico 2.12 a curva de rugosidade para o *Saccharomyces cerevisiae*. Neste gráfico foram apontados 4 valores de janela: $w = 71$, $w = 251$, $w = 501$ e $w = 905$. Os respectivos perfis de modularidade de janela podem ser vistos no gráfico 2.13.

É fácil perceber que a janela de largura $w = 71$ é o máximo da curva de rugosidade. Usando esta janela, teremos picos altos e vales profundos. Observe que os genes nas extremidades do ordenamento apresentam uma conectividade baixa (ver figura 2.13), portanto

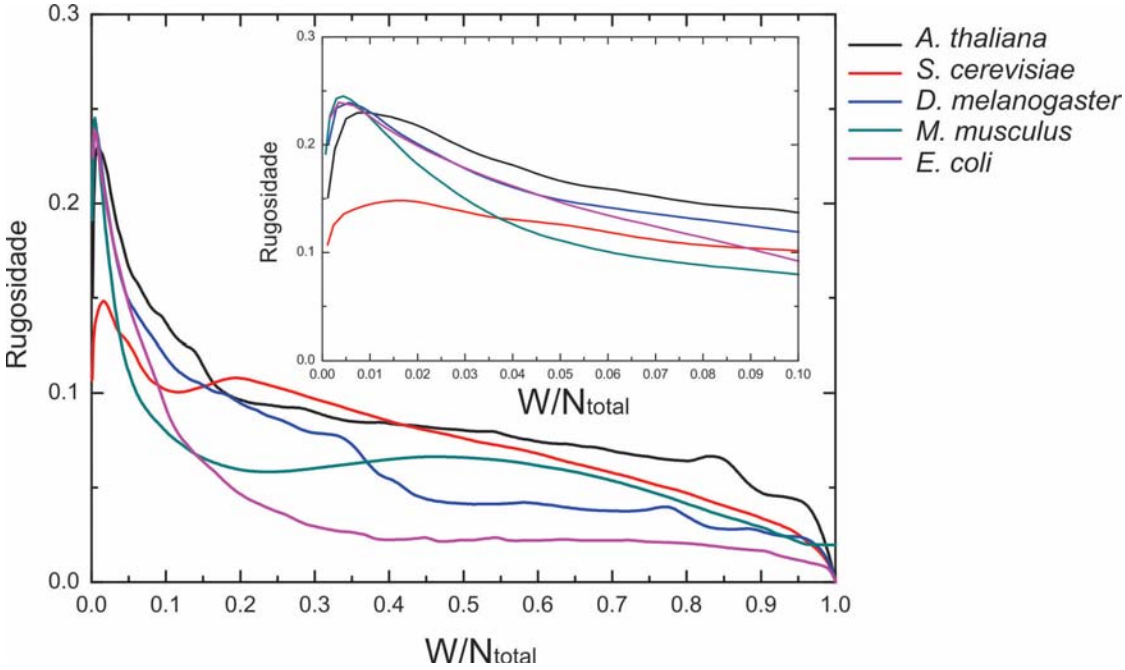


Fig. 2.11: *A figura apresenta as curvas de rugosidade para cinco organismos.*

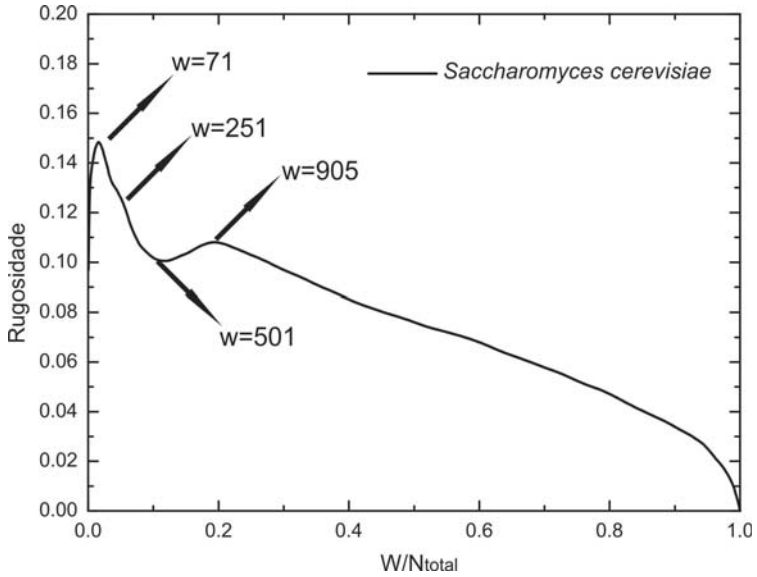


Fig. 2.12: *Curva de rugosidade para a Saccharomyces cerevisiae.*

a alta modularidade para estes genes não implica em grandes módulos. Esta janela consegue separar totalmente os genes poucos conectados na extremidade do ordenamento após a reorganização: a modularidade vai a zero. A janela de largura $w = 71$ mostra pequenos módulos. As larguras $w = 101$, $w = 151$ e $w = 201$ mostram uma diminuição na quantidade

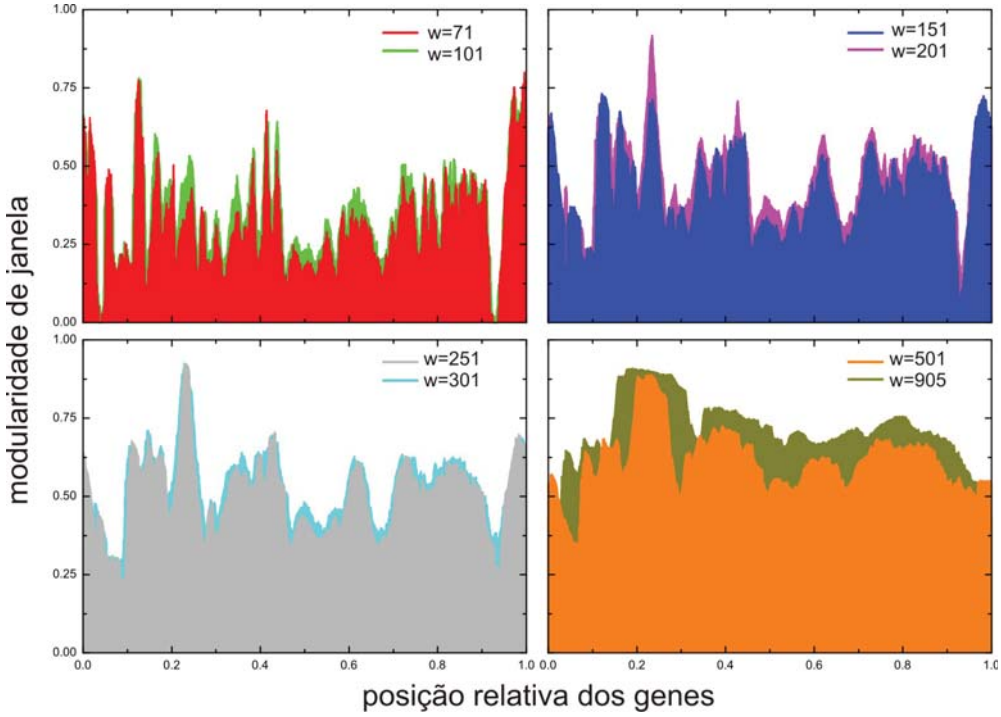


Fig. 2.13: *Perfis de modularidade de janela com largura(a)* $w = 71, 101$ *;(b)* $w = 151, 201$*; (c)* $w = 251, 301$*; (d)* $w = 501, 905$ *para Saccharomyces cerevisiae.*

de módulos, isso ocorre pelo crescimento de módulos. Já a janela de largura $w = 251$ é que será adotada em todo o desenvolvimento do trabalho; aparentemente ela mostra módulos de módulos. A largura $w = 301$ é muito similar a $w = 251$. As 501 e 905 janelas são muitos grandes e foram descartadas.

2.5 Comparação de Métodos

Além do método proposto para obter ordenação da rede, temos conhecimento de uma outra metodologia que pode ser usada para ordenar unidimensionalmente uma rede, que é método de *overlap* topológico, como proposto pelo Barabási e colaboradores e discutido na seção 1.3.3. Para compararmos os dois métodos, nós calculamos a densidade de pontos em cada diagonal da matriz de interação protéica após o ordenamento. Uma diagonal é representada pelos pontos $a_{i,i+n}$, onde n é um inteiro. Estes pontos representam interações entre genes localizados em posições separados por n na lista ordenada. Se a matriz tem tamanho $N \times N$,

a quantidade de diagonais superiores ou inferiores que ela possui é $N - 1$. A figura 2.14 mostra o estado inicial **(a)** e os respectivos estados finais para o método CFM **(b)** e o *overlap* topológico **(c)**.

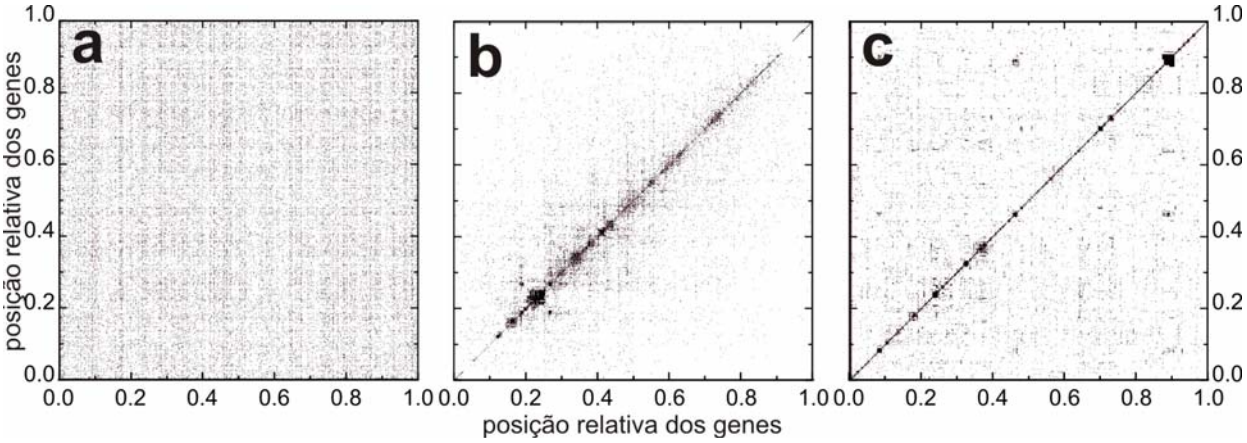


Fig. 2.14: *Grafico de possíveis ordenamento. (a) um ordenamento aleatório. (b) ordenamento final usando o método CFM. (c) ordenamento usando a métrica do overlap topológico*

Calculando a probabilidade de interação sobre as redes artificiais, obtemos a figura 2.15. Para rede **(a) e (b)** a probabilidade de interação decai como uma exponencial com o aumento da distância dos nós. E sabemos que a rede não forma nenhum módulo. Para a rede **(d)** o perfil da probabilidade de interação é constante. No caso da rede modular **(c)** a probabilidade é máxima para n menor que o tamanho do módulo, quando temos n igual à largura do módulo a probabilidade é minima. O método do *overlap* topológico consegue aglutinar mais pontos nas diagonais próxima à diagonal principal, porém ao longo da matriz essa densidade se mantém quase que constante. O CFM consegue fazer com que a densidade decaia exponencialmente à medida que n aumenta, isto é que a separação entre os genes cresce, como mostrado na figura 2.16 para *Saccharomyces cerevisiae*. O que torna o CFM melhor do que o método de *overlap* topológico para os nossos propósitos é consequência deste decaimento exponencial, como vamos discutir nas seções seguintes de análise funcional. Este decaimento exponencial mostra que a probabilidade que dois genes separados por n posições na lista ordenada interajam decai exponencialmente.

Suavizamos as curvas de conectividade e coeficiente de clusterização usando a mesma

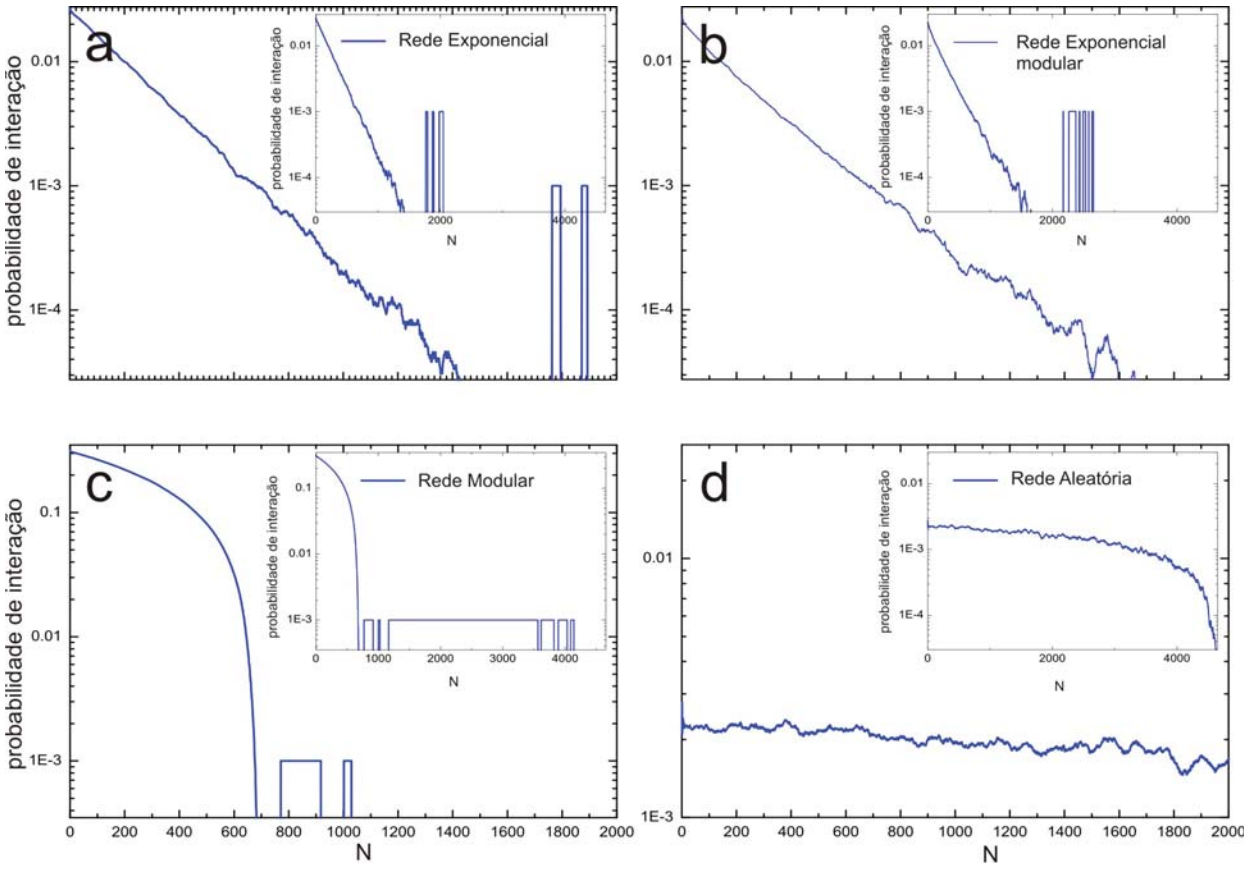


Fig. 2.15: *Painel da probabilidade de interação ao longo das diagonais das matrizes de interação. Gráfico maior é um zoom do gráfico menor. (a) rede exponencial, (b) rede exponencial modular, (c) rede modular e (d) rede aleatória.*

janela da modularidade: para o i-ésimo gene da lista calculamos a média da conectividade e do coeficiente de clusterização do intervalo de $w = 251$ genes centrado no i-ésimo gene. Estes perfis são mostrados nas figuras 2.17 para a *Saccharomyces cerevisiae*. Usamos condições de contorno periódicas. O método de *overlap* topológico apresenta mais picos que o CFM, devido ao fato de genes poucos conectados estarem mais espalhados pelo ordenamento,enquanto que o CFM deixa todos os genes poucos conectados nas extremidades. O pico da conectividade é maior no CFM e o coeficiente de clusterização está mais espalhado no ordenamento do *overlap* topológico do que no CFM, que apresenta um super módulo de alto coeficiente de clusterização no lado esquerdo do ordenamento.

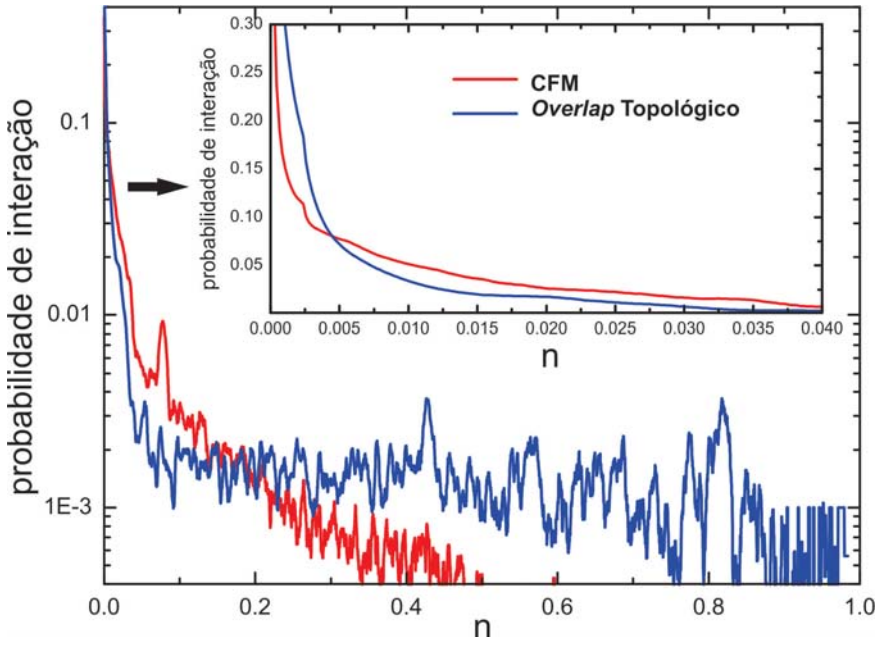


Fig. 2.16: *Gráfico da densidade de pontos das diagonais da matriz de interação versus n.*

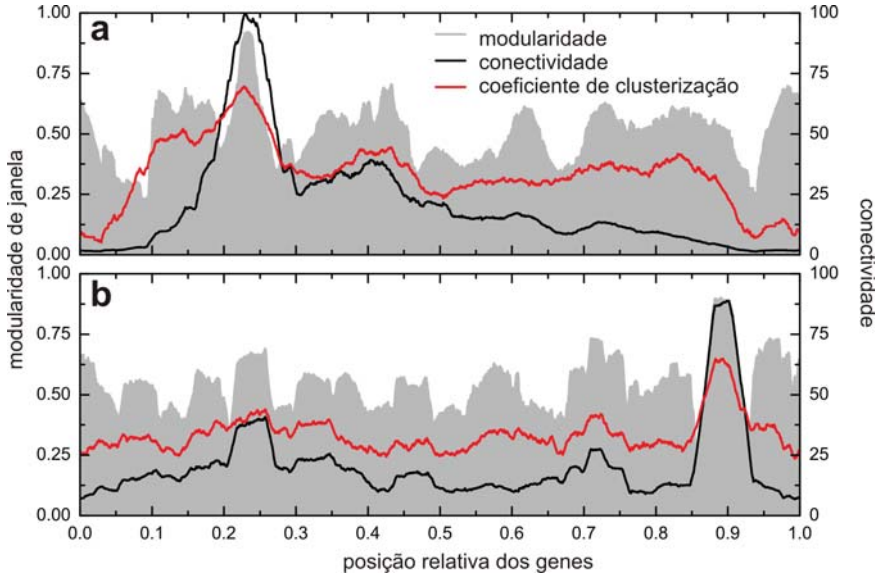


Fig. 2.17: *Acima temos o perfil de modularidade, juntamente com a conectividade e o coeficiente de clusterização obtido via CFM e abaixo o perfil gerado pelo método do overlap topológico.*

Capítulo 3

Análise Funcional

3.1 Enriquecimentos dos módulos

Como podemos descobrir quais processos biológicos estão inseridos nos módulos? Com o perfil de modularidade da rede, é possível identificar vários picos como mostrado na figura 2.17 (**a**). Neste caso o perfil de modularidade da *Saccharomyces cerevisiae* foi dividido em 7 picos. Cada pico é formado por um conjunto de genes e, com base nas interações obtidas com o STRING, foi construído uma rede para cada pico. As extremidades do perfil de modularidade não foram levadas em conta como um pico, pois temos genes pouco conectados presentes nestas duas regiões, devido à organização do algoritmo, nas extremidades sempre teremos genes com baixa conectividade.

No entanto, a identificação destes picos não informa qual o papel desempenhado por estes módulos no metabolismo celular. O *David Tools* auxilia este reconhecimento funcional.

Cada conjunto de genes representado por um pico no perfil de modularidade foi analisado usando *David Tools*. Entre as três ontologias do GO, selecionamos somente Processo Biológico para cada um dos 7 picos. Obtivemos assim 7 conjuntos de termos do GO (processos biológicos) associados a cada pico. Analisamos estes dados do ponto de vista bioquímico e selecionamos 33 termos como é apresentada na figura 3.1.

Estes dados foram plotados sobre o ordenamento da rede protéica, da seguinte maneira: foi feita uma comparação dos genes presentes nas tabelas do GO e no ordenamento. Construimos uma sequência binária $\phi(i)$ (onde, $i = 1, \ldots, N$) tal que, quando um gene de um determinado processo estivesse presente na tabela do GO e no ordenamento, localizado no

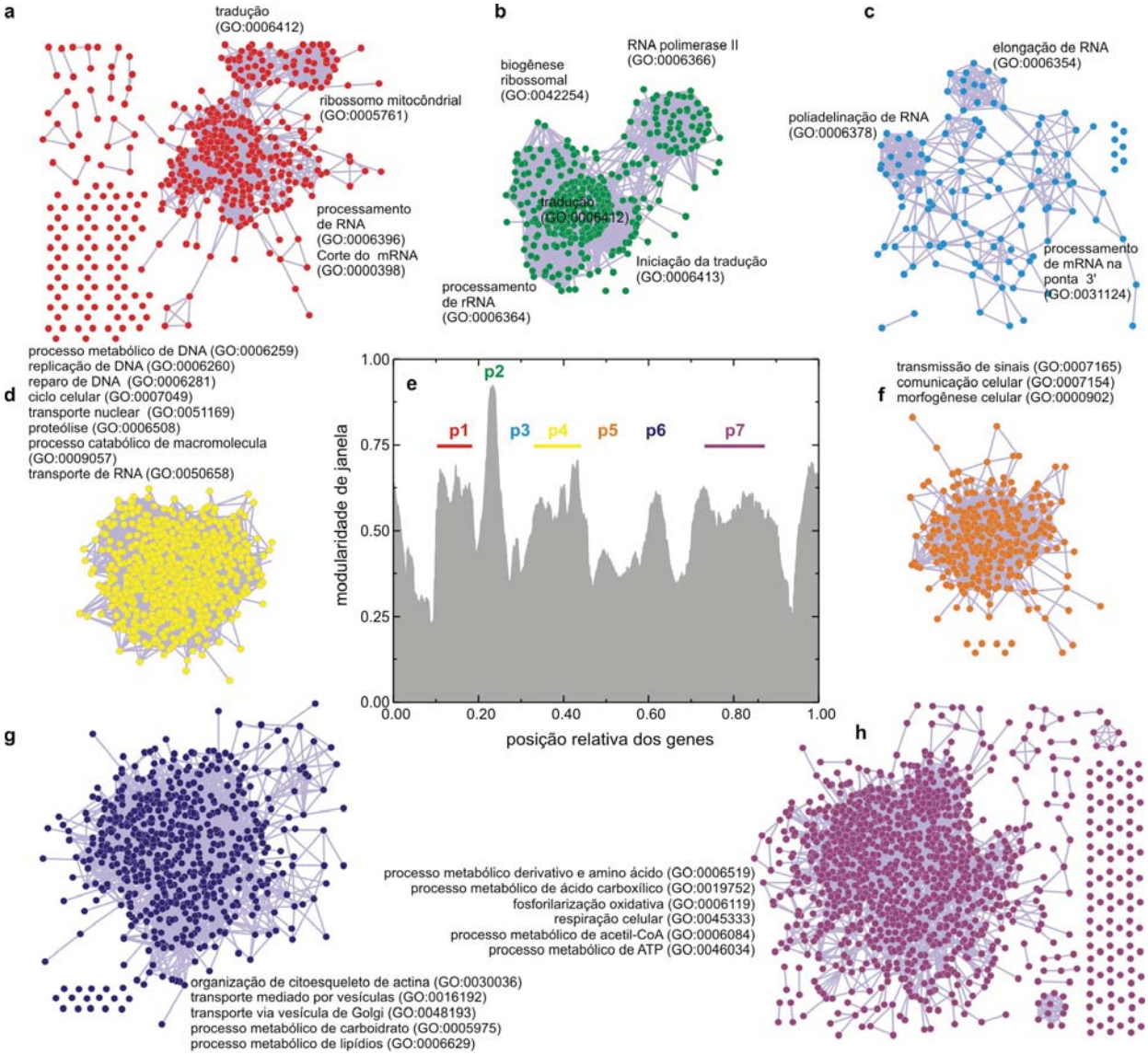


Fig. 3.1: *Painel dos módulos estudados e suas respectivas redes. Para cada uma das redes foram colocadas os nomes dos processos biológicos envolvidos*

ordenamento na posição i, atribuimos $\phi(i) = 1$. Quando o gene do ordenamento não estivesse presente na tabela da função, atribuimos $\phi(i) = 0$. Um gráfico simples desta sequência não traz informação devido às fortes flutuações. Faz-se necessário uma suavização desta pela média, o que dá lugar a uma densidade de valores entre 1 e 0 sobre a lista ordenada. Aqui ressaltamos que a relevância do ordenamento é a de possibilitar uma tal suavização pela definição da densidade de valores um sobre a sequencia: a proximidade de dois genes na lista correlaciona exponencialmente com a probabilidade de que estes dois genes interajam.

Este algoritmo de ordenamento é o único, que saibamos, capaz de dar lugar a uma lista com tal correlação entre posição na lista e probabilidade de interação. Para o cálculo da densidade $\mathcal{J}_w(i)$ consideramos janelas de $w = 251$ genes, a mesma usada para o cálculo da modularidade de janela. Para cada ponto central i da janela, atribuímos a média dos valores da sequência naquela janela, como definido na equação 3.1.

$$\mathcal{J}_w(i) = \frac{1}{w} \sum_{i-(w-1)/2}^{i+(w-1)/2} \phi(i) \tag{3.1}$$

Este processo é repetido para cada uma dos termos do GO apresentados na figura 3.1. O resultado são perfis que informam onde no ordenamento estes termos estão localizados. Picos destes perfis para uma dado termo significam que naquelas regiões concentram-se os genes associados àqueles termos.

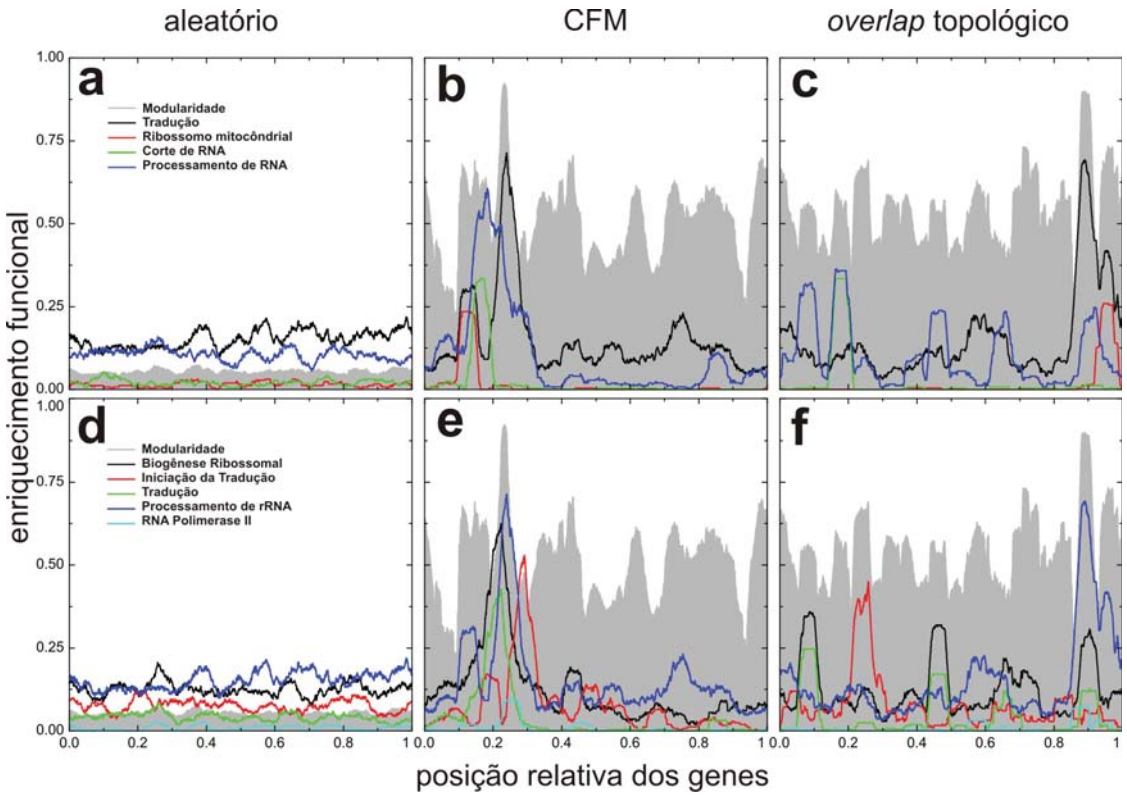


Fig. 3.2: *Perfis de termos do Gene Ontology: Processos Biológicos, projetados sobre o ordenamento aleatório, CFM e overlap topológico para os picos 1 (superior) e 2 (inferior). O perfil em cinza refere-se à modularidade.*

Ao plotarmos as funções sobre o ordenamento, observamos que tal organização ficou dividida em duas partes, os picos à esquerda são referentes a processos envolvendo produção de biomoléculas envolvendo portanto DNA e RNA (ver figura 3.1, itens **a**, **b** e **c**), enquanto que a parte direita refere-se ao metabolismo energético (ver figura 3.1, itens **d**, **f**, **g** e **h**). Esta característica é inerente à rede analisada.

As figuras 3.3, 3.4 e 3.5 são perfis de ordenamentos, aleatório, CFM e *overlap* associados aos processo biológicos definidos na figura 3.1.

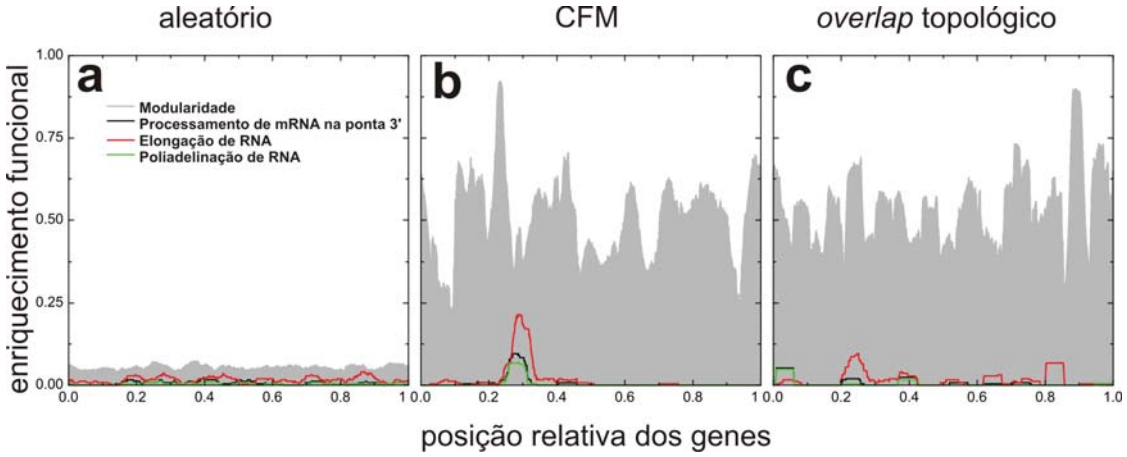


Fig. 3.3: *Perfis de termos do Gene Ontology: Processos Biológicos, projetados sobre o ordenamento aleatório, CFM e overlap topológico para o pico 3. O perfil em cinza refere-se à modularidade.*

Ao compararmos o ordenamento com o *overlap* topológico de Barabási, figura 3.2 **c** e **f**, podemos ver que muitas funções ficam espalhadas ao longo do ordenamento unidimensional. O CFM conseguiu concentrar o metabolismo de DNA, e o *overlap* topológico deixou este espalhado ao longo do ordenamento. Observe que os dois ordenamentos partem da mesma informação, aquela contida na matriz de adjancência. Ambas metodologias concentraram num ponto o processamento de RNA. E, na parte referente a metabolismo de energia, todas os processos estão espalhados na lista obtida com o método de *overlap* topológico, não havendo separação da parte energética da parte de DNA e RNA.

Observando a figura 3.4 **b**, vemos que o pico associado ao processo biológico de replicação de DNA sobrepõe-se ao de reparo de DNA e este sobrepõe-se ao pico associado a processos metabólicos de DNA. Do ponto de vista bioquímico, isto reflete o carácter hierárquico do

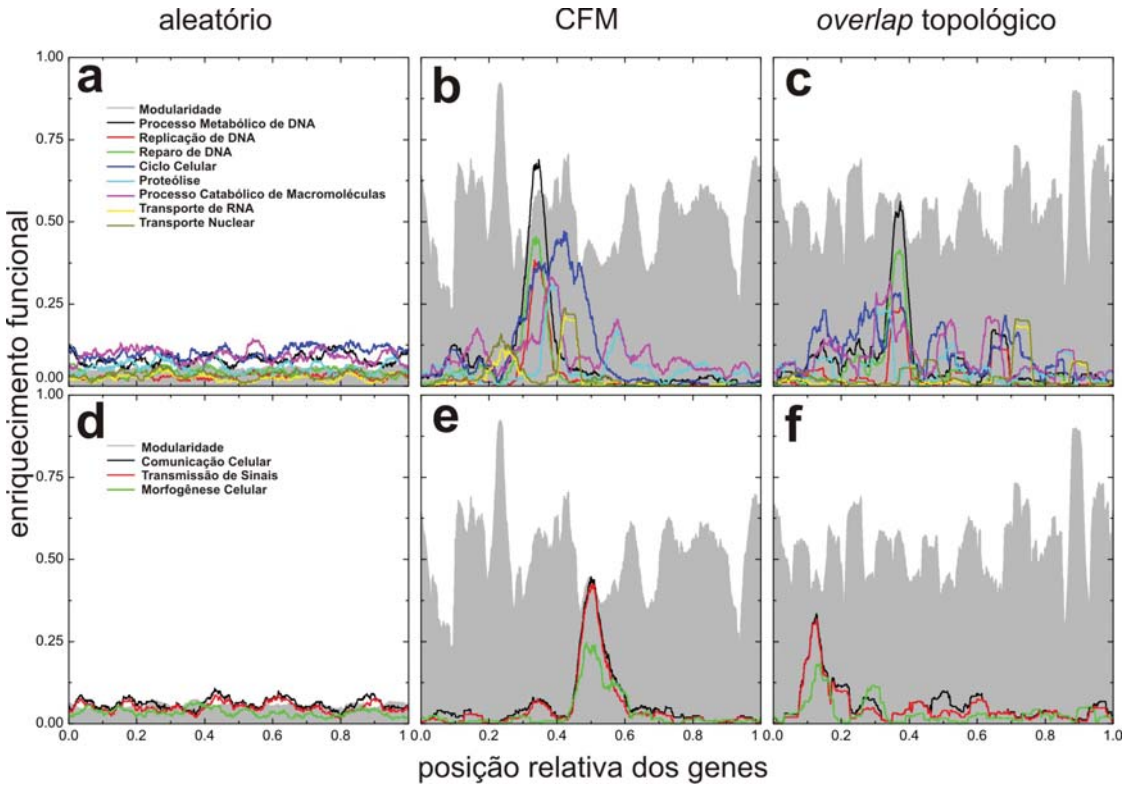


Fig. 3.4: *Perfis de termos do Gene Ontology: Processos Biológicos, projetados sobre o ordenamento aleatório, CFM e overlap topológico para os picos 4(superior) e 5 (inferior). O perfil em cinza refere-se à modularidade.*

GO: metabolismo de DNA é um processo que envolve vários outros processos onde o ácido desoxirribonucléico participa.

No caso temos os seguintes processos associados ao mesmo pico: reparo por excisão de bases, reparo por excisão de nucleotídeos, processos catabólicos, replicação, integração, modificação, proteção, recombinação, elongação, renaturação, mudança topológica, regulação negativa e positiva, DNA mitocondrial, manutenção de telômeros, etc.

A título de ilustração, nas figuras 3.2(**a**,**d**), 3.3(**a**), 3.4(**a**, **d**) e 3.5(**a**, **d**) plotamos os perfis associados para as ontologias usando um ordenamento aleatório. Como pode ser visto, as funções estão todas bem distribuídas e não há formação de picos. Não se pode assim obter informação relevante sobre a funcionalidade do sistema a partir de um ordenamento aleatório.

Com esta projeção dos termos do GO podemos associar os Processos Biológicos aos picos

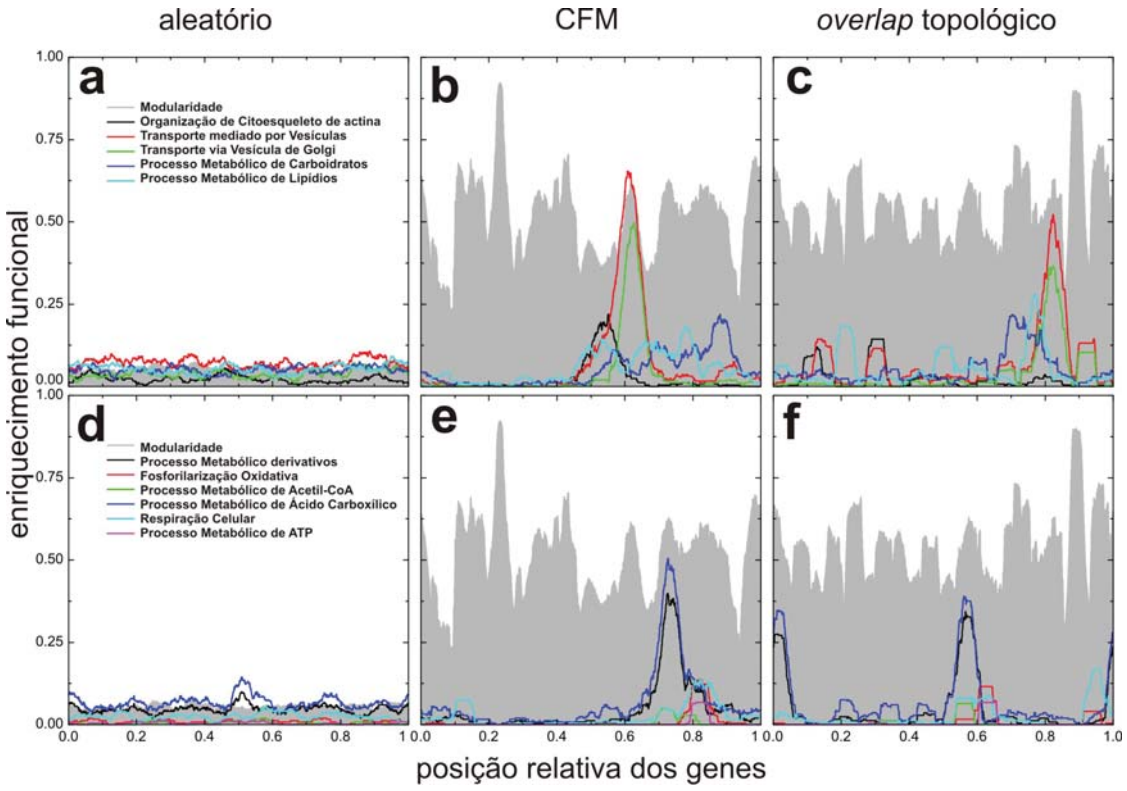


Fig. 3.5: *Perfis de termos do Gene Ontology: Processos Biológicos, projetados sobre o ordenamento aleatório, CFM e overlap topológico para os picos 6 (superior) e 7 (inferior). O perfil em cinza refere-se à modularidade.*

de modularidade, dando portanto um sentido biológico aos módulos identificados na matriz de adjacência após o ordenamento. Como veremos no próximo capítulo, este ordenamento é útil para medir a performance do metabolismo celular.

Capítulo 4

Aplicações

Neste trabalho expusemos brevemente o relato histórico sobre o DNA e o motivo de estudá-lo, propusemos uma maneira de organizar uma rede protéica em uma dimensão, apresentamos o conceito modularidade de janela como uma ferramenta para encontrar módulos funcionais e mostramos uma maneira de associar um significado biológico a cada módulo. Todos estes passos foram necessários para desenvolver um método para análise de expressão gênica de genoma completo, que aqui será chamado de **Transcriptograma**.

4.1 Transcriptograma

Hoje em dia qualquer análise de expressão gênica é realizado com poucos genes. Por exemplo, em um experimento que compara uma cultura de células cancerosas com células normais, depois de obter os dados através de um microarranjo, o pesquisador deve escolher quais genes devem ser levados em conta na análise e quais devem ser descartados.

Na maioria dos casos o pesquisador analisa os níveis de expressão de cada gene e estabelece um limiar. Por exemplo, o pesquisador pode considerar somente os genes que tiveram níveis duas vezes superior e duas vezes inferior em relação ao normal. Com este conjunto de genes superexpressos e subexpressos o pesquisador conduz sua análise. Geralmente o conjunto de genes não é superior a 50, sendo que um microarranjo de *Saccharomyces cerevisiae* tem cerca de 6000 genes, de *Homo sapiens* tem mais de 30 mil genes.

Observamos que o genoma funciona na forma de uma rede e, sendo assim, um módulo de muitos genes medianamente superexpresso ou subexpressos podem causar um efeito não desprezível. No entanto tal modificação poderia não ter sido analisada seguindo o critério

de escolher genes com modificação no nível de expressão muito intensas.

Uma maneira de contornar essa situação é usar um método que considere os níveis de expressão do genoma completo, o que é contemplado pelo transcriptograma.

4.2 Ciclo celular da *S. cerevisiae*

Na referência [TU 05] Tu e colaboradores cultivaram a levedura até cobrir toda a (confluência) placa de cultivo, privaram as células de alimento e depois de 10 horas liberaram glicose. Com isto as leveduras entraram em fase, apresentando ciclos respiratórios sincronizados. Cada ciclo é caracterizado por uma oscilação nos níveis de O_2 dissolvido no meio de cultura ao alternarem-se as fases de fermentação e fase oxidativa (respiração). Cada ciclo leva em torno de 5 horas. No instante $t = 0$ foi retirada uma amostra, com a qual foi realizado um experimento de microarranjo, onde medem-se os níveis de expressão dos genes da *Saccharomyces cerevisiae*. A cada intervalo de 25 minutos foi retirada uma nova amostra para medir os níveis de expressão. Foram considerados 3 ciclos de fermentação-respiração, totalizando 36 transcriptomas de genoma completo. Os dados dos transcriptomas deste experimento estão disponíveis no GEO sob o código *GSE3431*. Ao perceberem que muitos genes apresentavam oscilações no níveis de expressão ao longo do tempo, os pesquisadores realizaram uma análise de clusterização usando o método *k-means* [MCL 08] com todo os dados contidos nos microarranjos, e agruparam os genes em 3 grupos distintos. Foram estudados somente 120 genes (40 em cada grupo), pois apresentaram as melhores correlações na clusterização. Os grupos receberam os nomes de *oxidative*, *reduction/charging* e *reduction/building*.

- *Oxidative:* nesta fase existe o consumo de oxigênio e os genes que participam nesta etapa são aqueles que codificam proteínas ribossomais, síntese de ribossomos, síntese de aminoácidos e metabolismo de enxofre.

- *Reductive/building:* é quando as células começam a cessar o consumo de oxigênio. A maioria dos genes codificam proteínas mitocondriais, replicação de DNA, divisão celular, etc.

- *Reductive/charging:* está relacionado com a codificação de proteínas que envolvem modos não-respiratórios do metabolismo e degradação de proteínas, tais como: glicólise, oxidação de ácido gorduroso, proteassoma, ubiquitinação, etc.

Tomamos cada um dos transcriptomas e produzimos perfis de expressão sobre o ordenamento, sempre considerando médias sobre janelas de $w = 251$.

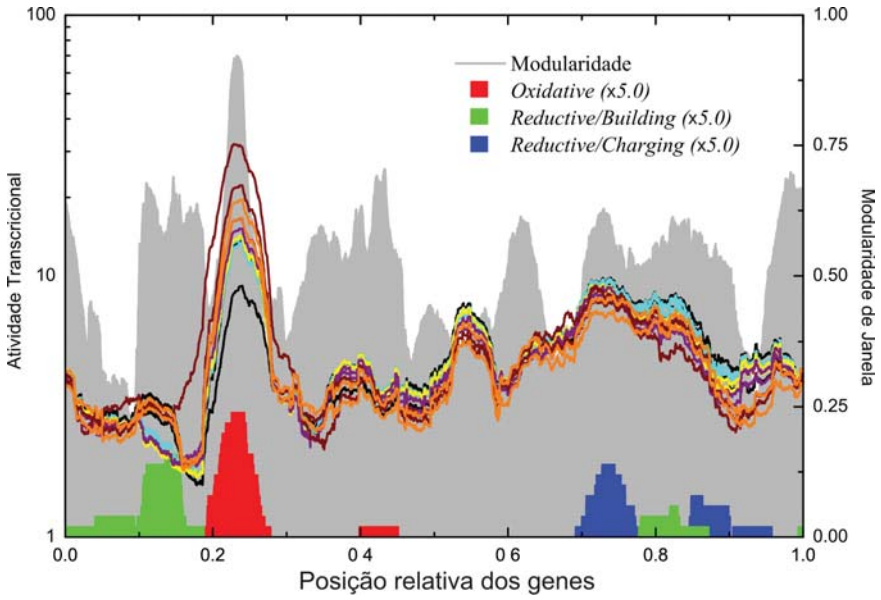


Fig. 4.1: *Transcriptograma do ciclo metabólico da levedura sobre ordenamento do CFM.*

A figura 4.1 mostra o perfil de modularidade ao fundo, juntamente com o perfil dos 3 grupos com 40 genes cada, que foram estudados no trabalho de Tu *et al.* [TU 05]. Para melhor visualização na figura, multiplicamos a densidade de janela destes 3 grupos por um fator 5. As posições destes picos no ordenamento estão de acordo com a análise funcional vista no capítulo anterior. Na mesma figura é adicionado o perfil de atividade transcricional do primeiro ciclo de respiração-fermentação (12 transcriptomas).

Podemos ver que existe um *gap* no ordenamento que foi deixado de lado. Agora vamos nos fixar entre os valores 0.35 e 0.7. Vemos um pico de expressão entre 0.35 e 0.45, este pico apresenta um comportamento oscilatório. Buscamos os 40 genes com maiores níveis de expressão entre a posição 0.35 e 0.45 no $t = 25$min, pois é onde ocorre o máximo.

Para uma análise estatística, consideramos cada ciclo (12 transcriptomas) como um conjunto de replicatas. Para cada gene i dos três perfis de expressão apresentados em cada

quadro foi calculado o desvio padrão amostral (equação 4.1) do conjunto. Os termos $e_w(i)$ e $\bar{e}_w(i)$ são respectivamente: a expressão do gene i para uma janela w e a expressão média ($t = 0$ min, $t = 300$ min e $t = 600$ min) do gene i com janela w. No fundo de cada

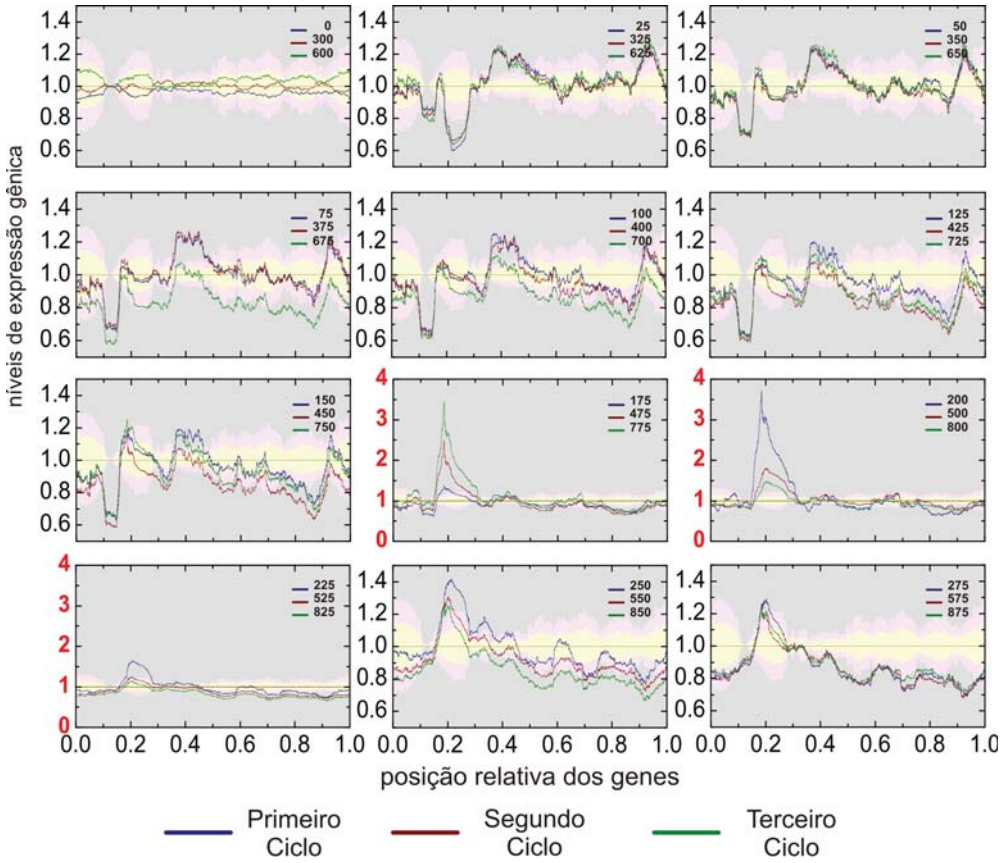


Fig. 4.2: *Transcriptograma relativo do ciclo metabólico da levedura usando o ordenamento obtido pelo CFM para uma janela de* $w = 251$*. A banda amarela corresponde de* $1 - 2\sigma$ *a* $1 + 2\sigma$*, a banda rosa corresponde de* $2\sigma - 4\sigma$ *a* $2\sigma + 4\sigma$ *e a banda cinza correponde a desvios maiores do que* 4σ

gráfico (figura 4.2) temos três faixas de cores, o amarelo representa desvio de 0 à 2σ, o rosa corresponde de 2σ à 4σ e cinza é maior do que 4σ. Esta figura revela a significância estatística da superexpressão e subexpressão dos grupos R/B, R/C, Ox e o pico entre 0.35 e 0.45.

$$\sigma_w(i) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(1 - \frac{e_w(i)}{\bar{e}_w(i)})^2} \tag{4.1}$$

Acompanhando a evolução temporal do ciclo de respiração-fermentação, vemos que os picos indicados por Tu *et al.* estão quatro desvios padrões acima do normal, e o pico entre 0.35 e 0.45 também se encontra acima de quatro desvios.

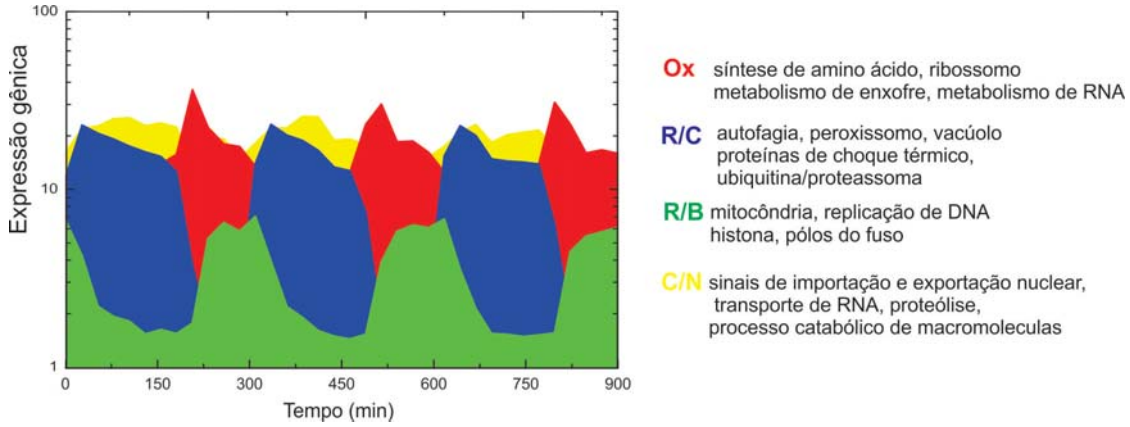


Fig. 4.3: *Atividade transcricional média de cada grupo estudado por Tu et. e o grupo (amarelo) descoberto pelo método CFM.*

A figura 4.3 apresenta a atividade transcricional média para cada tempo t de cada um dos grupos estudados por Tu *et al.*, e colocamos o grupo de genes do intervalo 0.35-0.45. Existe um comportamento oscilatório dos 4 conjuntos, o conjunto C/N (processos catabólicos e transporte nuclear) complementa os outros 3. Este grupo de genes está relacionado à quebra de moléculas como a glicose ou triglicerídeos e transformam-nas em moléculas menores pobres em energia. A energia liberada é armazenada em forma de ATP e GTP que são necessárias para a importação ou exportação de macromoléculas para o núcleo da célula.

As figuras 4.4, 4.5 são referentes respectivamente ao ordenamento CFM e ordenamento *overlap* topológico. Cada curva representa um sinal de microarranjo que foi dividido pelo sinal médio dos microarranjos iniciais de cada ciclo ($t = 0$ min, $t = 300$ min e $t = 600$ min). Ao todo foram realizados 36 medidas dos trancriptomas. Observando a figura relativa ao ordenamento CFM vemos que reproduz os dados do artigo [TU 05]. No intervalo de 25-50 minutos temos uma baixa expressão relativa nos módulos referentes a processo de RNA, tradução. A parte referente a metabolismo está super-expressando o que significa que a levedura está fermentando (há produção intensa de biomoléculas o que significa consumo alto de energia). Entre 50-175 minutos, há uma lenta redução da produção de biomoléculas

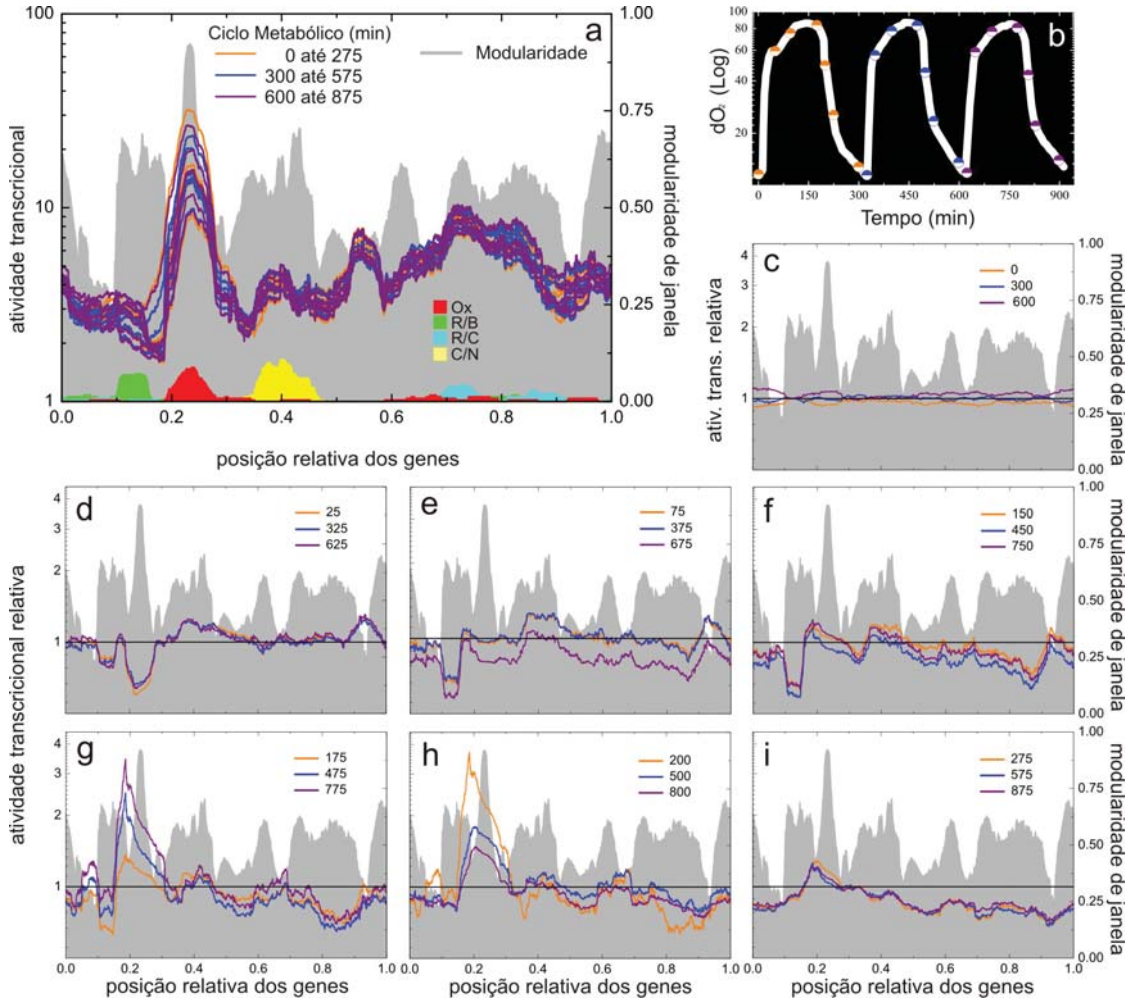


Fig. 4.4: *Transcriptograma relativo do ciclo metabólico da levedura usando o ordenamento obtido pelo CFM, com janela* $w = 251$.

e o oxigênio inicia lentamente o processo de quebra de biomoléculas para produção de energia. No intervalo de 175-200 minutos temos um máximo que é referente ao pico de oxidação que está na figura 4.4 (**a**) (pico em vermelho). No ordenamento gerado pelo critério de Barabási, figura 4.5, este pico também é visto, mas dividido em outros 3 picos, um em cada extremidade do ordenamento e um no centro. De 175-275 minutos, o grupo *reductive/charging* apresenta a sua expressão aumentada em relação ao ($t = 0$), isto indica que inicia o processo fermentativo e se estende até o $t = 275$ min.

A título de ilustração, apresentamos um transcriptograma gerado por um ordenamento aleatório (sem correlação), figura 4.6, fica nítido que não é possivel distinguir nenhuma

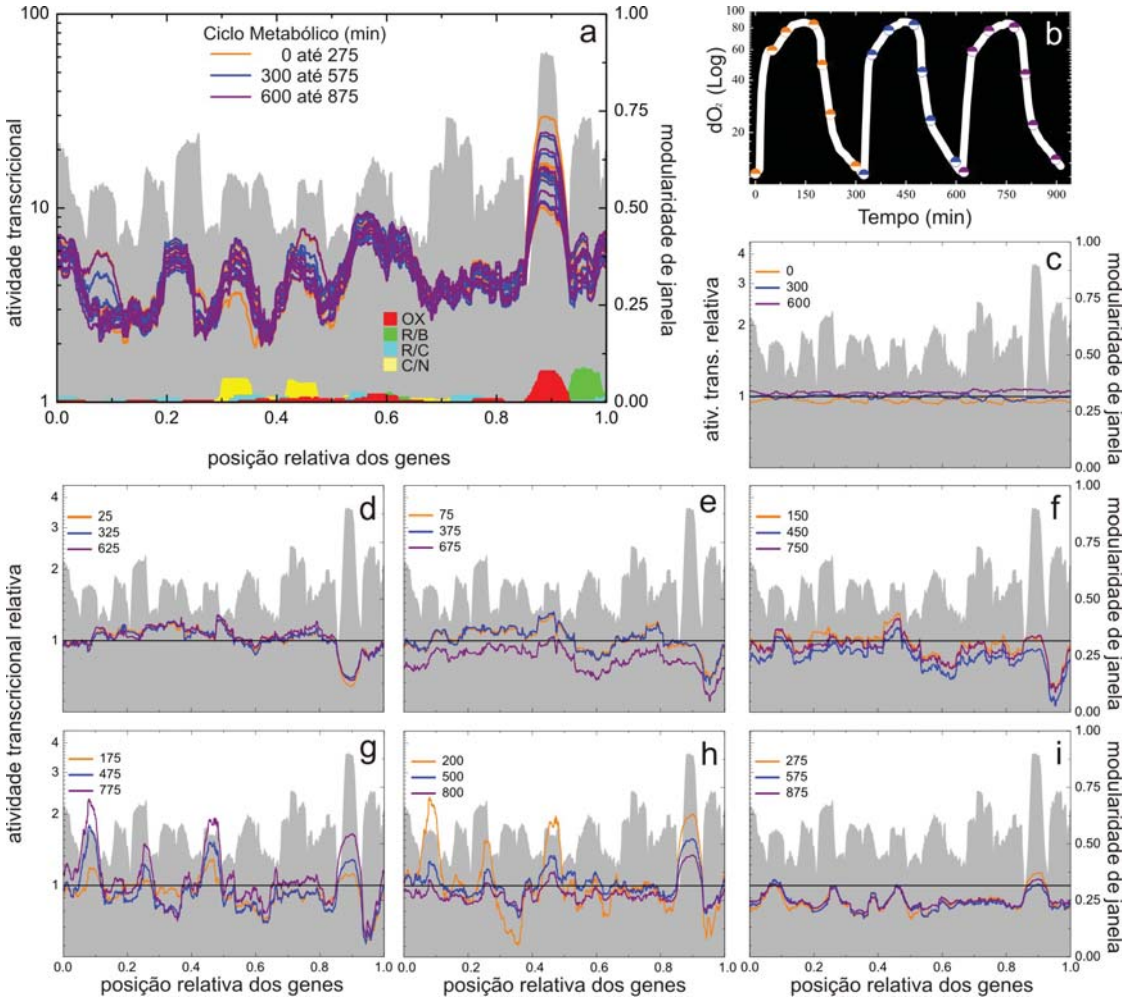


Fig. 4.5: *Transcriptograma relativo do ciclo metabólico da levedura usando o ordenamento obtido pelo overlap topológico, com janela* $w = 251$.

ativação ou desativação de módulo funcional, exceto que o nível médio de expressão relativa oscila com o tempo.

4.3 Célula normal versus célula danificada

No trabalho de Fry *et al.* [FRY 03] foi estudado a função gene SGS1 da *Saccharomyces cerevisiae* que possui dois genes homólogos (BLM e WRN) em *Homo sapiens*. Tais genes estão relacionados à alopécia, predisposição ao câncer, catarata, perda da resilência da pele, etc. Deletar o SGS1 do DNA causa o envelhecimento precoce da levedura com redução do

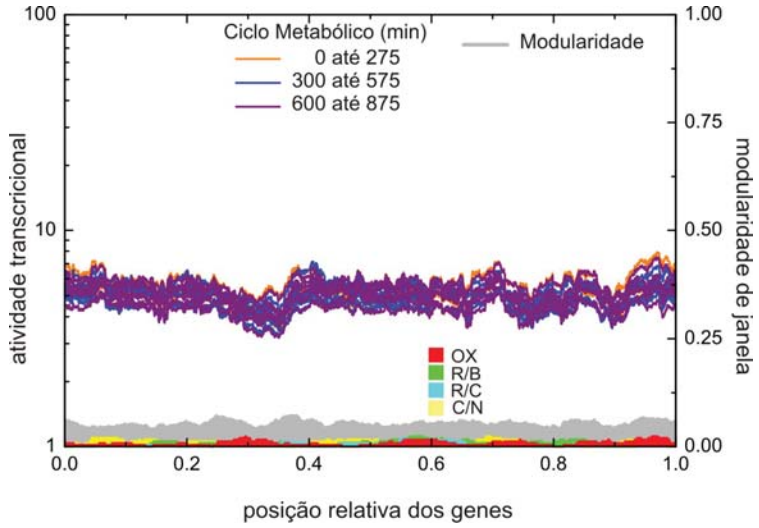


Fig. 4.6: *Transcriptograma relativo do ciclo metabólico da levedura usando o ordenamento aleatório, com janela* $w = 251$.

tempo de vida em cerca de 40%.

No caso de SGS1 Fry *et al.* consideraram 188 genes, detectados como aqueles que apresentaram as maiores alterações nos níveis de expressão de mutantes com SGS1 deletado em relação à linhagem normal, veja figura 4.7. O SGS1, no entanto, é um gene regulador e uma simples alteração neste gene pode refletir uma alteração em quase todo o genoma. Muitos efeitos podem passar despercebidos numa seleção de poucos genes. Neste estudo Fry *et al.* consideram que apenas 3-4% do genoma sofreu alteração. Os pesquisadores só consideraram genes alterados quando a sua expressão gênica era duas vezes superior ou inferior em relação à expressão do gene normal. Os dados dos transcriptomas deste experimento estão disponíveis no GEO sob o código *GSE423*.

As cepas deste estudo são células com o SGS1 deletado (caso) e uma linhagem normal (controle). Para separar as leveduras que possuiam SGS1 deletado das normais foi usada Geneticina com uma concentração de $20\mu g/ml$. A levedura cresceu num meio de YEPD (*Yeast-Extract Peptone Dextrose*) que contém 1% de extrato de levedura, 1% de peptona e 1% de glicose. Ela foi mantida à temperatura de 30°C. Cada uma delas foi dividida em duas. Uma amostra de SGS1 deletado recebeu 0.1% de metil-metanosulfonato (MMS) que foi adicionado diretamente sobre a amostra. Uma amostra da linhagem normal também

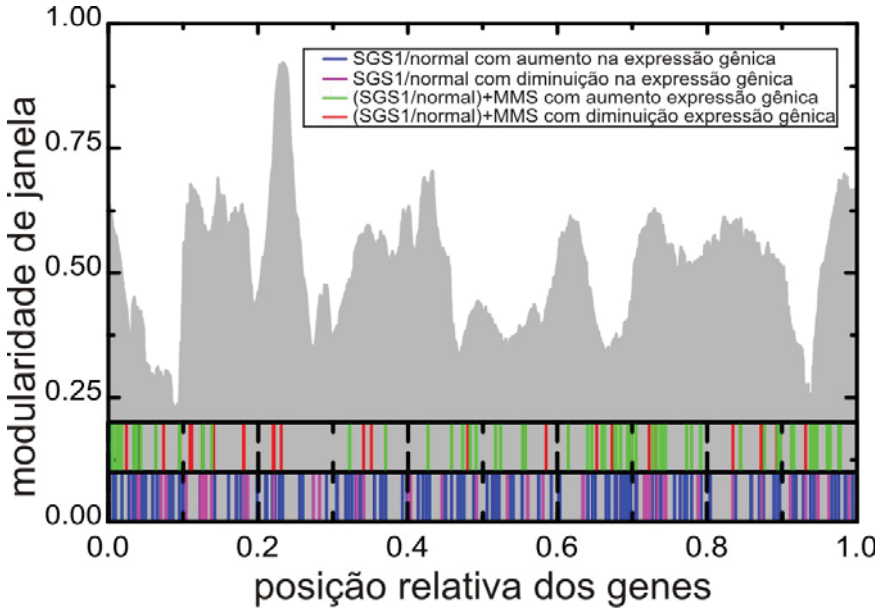


Fig. 4.7: *O perfil de modularidade da Saccharomyces cerevisiae e abaixo é apresentado a posição dos genes estudadas por Fry et al.. Para a célula com SGS1 deletada em comparação com a célula normal, nós temos o azul que corresponde ao aumento da expressão gênica e o púrpura que indica a diminuição aa expressão. Para as células que tiveram a aplicação de MMS, o verde refere-se aos genes que tiveram aumento na expressão e o vermelho indica a diminuição da expressão.*

recebeu 0.1% de MMS. Após a aplicação do MMS, as cepas ficaram incubadas durante 1h à temperatura de 30ºC. O MMS foi retirado com uma lavagem com água destilada e com o uso de TES Buffer (10 mM de Tris-hidroximetil-aminometano (Tris) , 10mM de ácido etilenodiamino tetra-acético (EDTA) e 0.5% de Dodecil sulfato de Sódio (SDS)). Após estes tratamentos foram realizadas as medidas com uso de microarranjo.

Podemos perceber na figura 4.8 que tanto a cepa normal quanto a cepa com o gene SGS1 deletado ambas com o MMS, possuem o mesmo perfil de expressão. O MMS causa o mesmo tipo de dano em ambas amostras. Por outro lado o SGS1 está relacionado com genes que participam da replicação de DNA e reparo de DNA, ao deletá-lo a célula tem uma subexpressão na região referente à replicação e reparo, que é entre 0.3-0.4 no eixo das abscissas. Existe dano na morfologia celular, na comunicação celular, processo metabólico de carbohidrato,etc. Também observamos que a expressão relativa das linhagens com SGS1

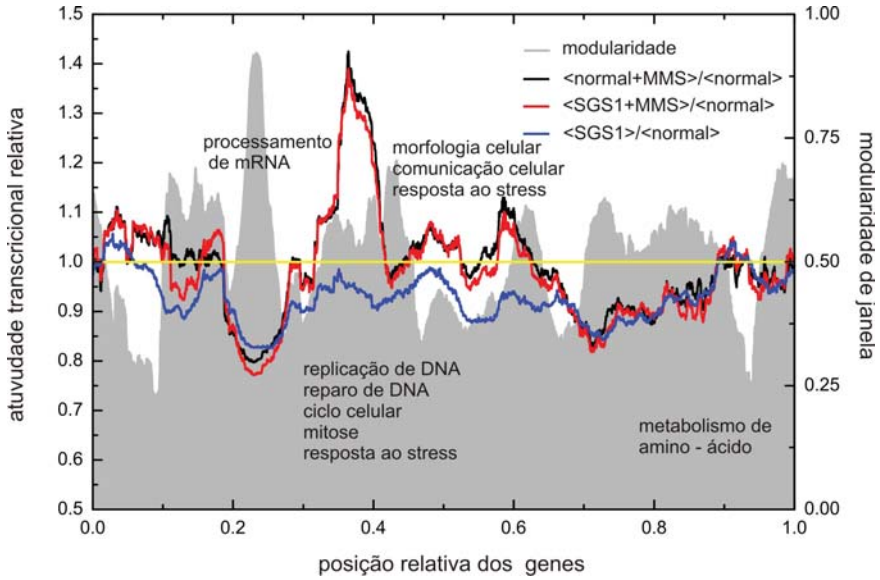


Fig. 4.8: *Transcriptograma da levedura plotado sobre o ordenamento do CFM, neste gráfico temos todos os perfis de transcrição médios divididos pela media da transcrição da amostra normal (atividade transcricional relativa). A reta amarela corresponde a médias dos transcriptomas normais sem a adição de MMS.*

deletado em relação ao normal está subexpressado ao longo de todo o genoma. No entanto os autores consideram somente de 3%-4% do genoma como alterado. Observando que se o SGS1 é um gene regulador, é razoável que a uma simples alteração nele possa causar uma alteração em todo o transcriptoma e não somente em uma pequena porção.

O transcriptograma é um método de análise de dados de micro-arranjos que considera a expressão de todos os genes. Os módulos de genes medianamente superexpresso ou subexpressos podem ter uma importância fundamental no metabolismo da célula e não devem ser desconsiderados *a priori*.

A exposição de *Saccharomyces cerevisiae* a MMS tem a propriedade de parar o ciclo celular em uma das três possíveis fases, (**G1**, **S** e **G2**) [JEL 00]. Essas fases pertencem ao processo pelo qual as células eucarióticas dividem os seus cromossos e dão origem a duas células-filhas (**mitose**). O ciclo celular possui dois estágios maiores, a interfase e a fase mitótica. A interfase compreende a fase **G1** que é caracterizada por expressão de genes e síntese de proteínas. Isto permite à célula crescer e produzir todas as proteínas necessárias para a síntese de DNA. Em seguida temos a fase **S**, quando a célula replica seu DNA. Essa

replicação deve ser precisa a para prevenir anormalidades genéticas que podem levar a morte da célula ou doenças genéticas. Isto permite à célula dividir-se em duas células filhas, cada uma delas com uma cópia completa do DNA da célula mãe. A levedura é um organismo modelo para o estudo da fase **S**, pois as rotas metabólicas que governam este processo são altamente conservadas. E na fase **G2**, a célula novamente cresce e aumenta a síntese de proteínas e do RNA, fazendo com que o gasto de energia cresça, e assim inicia a duplicação dos centríolos que irá permitir o processo de divisão.

Na figura 4.9 apresentamos o transcriptograma para cada uma das duas replicatas das amostras de células normais e das células mutadas sem a adição de MMS. Os níveis de expressão estão divididos pela média aritmética dos transcriptogramas de célula normal. As curvas laranja e vermelha mostram que as células com gene SGS1 deletado têm todo o seu metabolismo diminuído em relação a células normais.

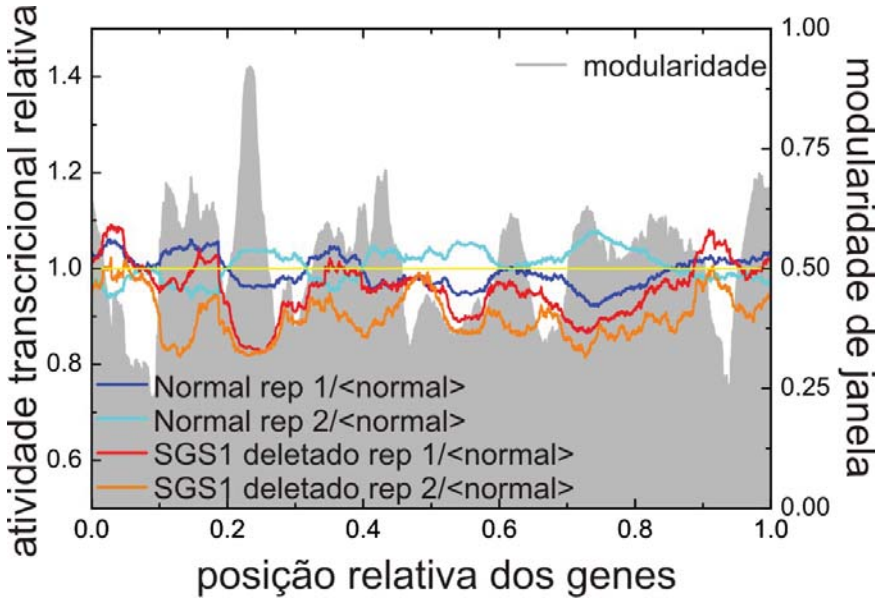


Fig. 4.9: *Transcriptograma da levedura plotado sobre o ordenamento do CFM, neste gráfico temos todos os perfis de transcrição divididos pela média dos perfis normais, sem adição de MMS. A reta amarela corresponde a média dos transcritomas normais sem adição de MMS*

Na figura 4.10 (**a** e **b**) temos o transcriptograma para as células normais e danificadas com a adição do MMS. Existe diferença na expressão gênica entre as replicatas. O módulo

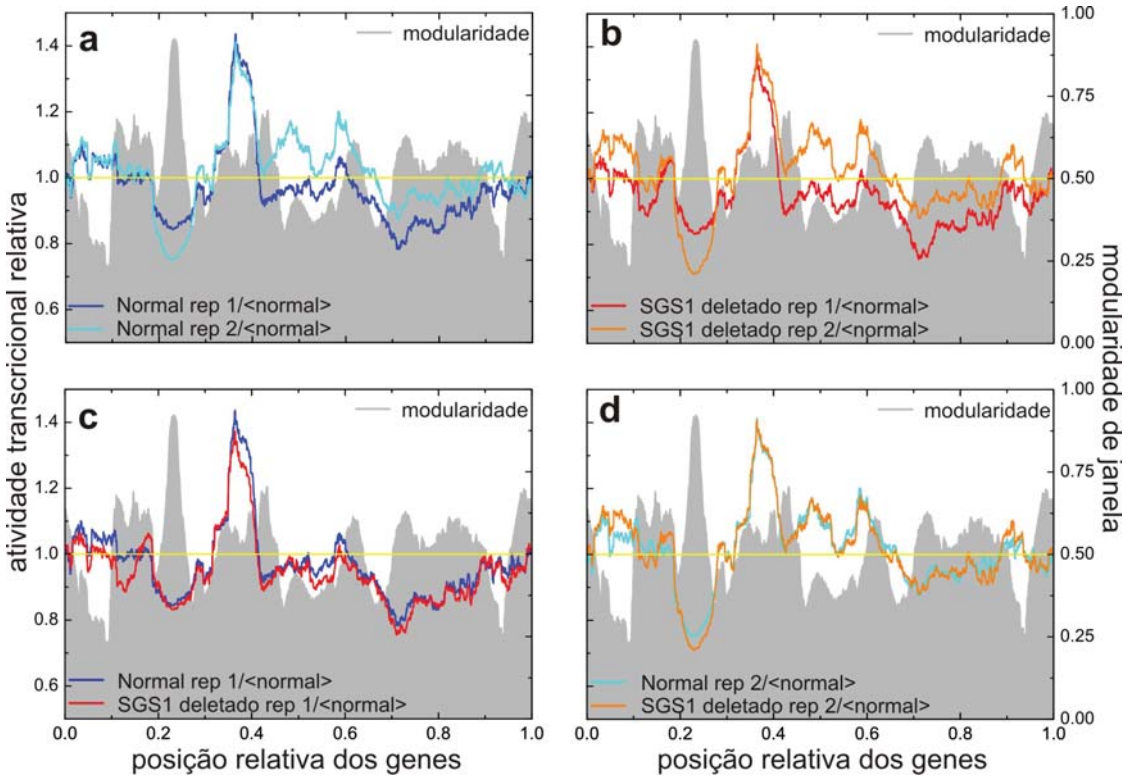


Fig. 4.10: *Transcriptograma da levedura plotado sobre o ordenamento do CFM, neste gráfico temos todos os perfis de transcrição divididos pela média dos perfis normais, sem adição de MMS. A reta amarela corresponde a média dos transcritomas normais sem adição de MMS*

responsável pela tradução tem um baixo nível de expressão para as replicatas número 2 do que para as replicatas número 1, tanto para a célula normal quanto para a danificada. Os módulos de comunicação celular, morfogênese celular, processo metabólico de lipídios e respiração celular estão mais superexpressos nas replicatas número 2 do que nas replicatas número 1.

Agrupando o trancriptograma da célula normal e danificada das replicatas rotuladas pelo número 1, figura 4.10 (**c**), e fazendo o mesmo para as replicatas número 2 (figura 4.9 (**d**), podemos notar que em cada gráfico as replicatas apresentam transcriptogramas equivalentes, mas as amostras da figura 4.9 (**d**) estão em fases diferentes daquelas da figura 4.10 (**c**). Fry e seus colaboradores não perceberam em suas análises o fato que de que as replicatas estavam em fases diferentes. O método apresentado mostrou-se capaz de diferenciar estágios do ciclo

celular (respiração-fermentação) e diferenças entre células normais e mutadas. Esses fatos garantem que o método tem poder de diagnóstico e o próximo passo será aplicar em tecidos cancerosos.

Capítulo 5

GNATT

A análise dos dados apresentada nas seções anteriores requer diferentes etapas e habilidades. Primeiramente faz-se necessário a busca dos dados em diferentes locais, que se utilizam de diferentes nomenclatura. Estes dados estão organizados de forma diferente, segundo diferentes critérios embasados em propriedades bioquímicas dos genes, seus produtos e interações.

A escolha de como lidar com essas classificações requer *expertise* bioquímica a qual tivemos acesso pela intensa e frutífera interação com nossos colaboradores do departamento de Bioquímica da UFRGS. A maneira de agrupar dados com vistas à definição de módulos funcionais, por outro lado, requer *expertise* na análise e proposição de modelos de sistemas complexos, um campo no qual nosso grupo de pesquisa vem trabalhando há vários anos. Durante a execução do presente projeto, bem como de outros correlatos, fomos encontrando tarefas que exigem a presença simultânea de físicos e bioquímicos. Tais tarefas podem se subdivididas em dois grupos. Aquelas onde os critérios, hipóteses ou resultados são discutidos, comparados e questionados, e um segundo grupo, que tratam da organização dos dados segundo alguma hipótese de trabalho. Para este segundo grupo, é possível operacionalizar as tarefas necessárias por meio de aplicativos computacionais que facilitam sua realização. Para isso, este projeto contempla a criação e disponibilização em rede do GNATT, um aplicativo que possibilita a manipulação de dados para, entre outras, a produção de transcriptogramas.

Durante o desenvolvimento do GNATT foi construído o VIACOMPLEX [CAS 09]. O VIACOMPLEX é uma ferramenta computacional que constrói mapas de expressão gênica sobre uma rede de interação previamente clusterizada pelos programas PAJEK [dN 05] ou MEDUSA [HOO 05], calcula o *overlap* topológico da rede e estima os níveis de expressão

relativa de grupos de genes funcionalmente associados. A diferença entre o GNATT e o VIACOMPLEX, é que o primeiro trabalha com a minimização da função custo aplica a redes e o segundo utiliza redes previamente processadas pelos programas PAJEK [dN 05] ou MEDUSA [HOO 05].

O GNATT é uma ferramenta computacional de análise de dados de interação (disponível no endereço *http://lief.if.ufrgs.br/pub/biosoftwares/Gnatt/*), onde podemos reorganizar uma rede via os dois métodos discutidos nas seções anteriores: Minimização da Função Custo e o *overlap* topológico. Depois destes dados organizados, pode-se proceder à análise funcional. A página inicial da ferramenta apresenta um painel de análise funcional onde é possível calcular a modularidade de janela e identificar seus picos e então proceder com uma análise de expressão gênica projetada sobre o ordenamento unidimensional. O objetivo do GNATT é que qualquer usuário possa reproduzir a análise apresentada neste trabalho, utilizando diferentes bancos de dados públicos.

A ferramenta foi construída com Intel Fortran, o software possui seis funções básicas explícitas, como segue: Análise de clusterização, Análise Funcional, Análise de Expressão, Criação de Interatoma, Busca por camadas, Análise Estatística. O software GNATT encontra-se em versão beta e está disponível na internet.

5.1 Análise de Clusterização

Esta subrotina possui dois métodos de reorganização de dados de interação protéica (figura 5.1) que foram apresentados no capítulo 1 e no capítulo 3:

- *Overlap* Topológico
- CFM

No caso do CFM existe uma série de parâmetros que devem ser inicializados, como segue:

- **Exponent:** é o expoente α ao qual elevamos a distância calculada entre um elemento não nulo e a diagonal da matriz de adjacência, como apresentada na expressão da

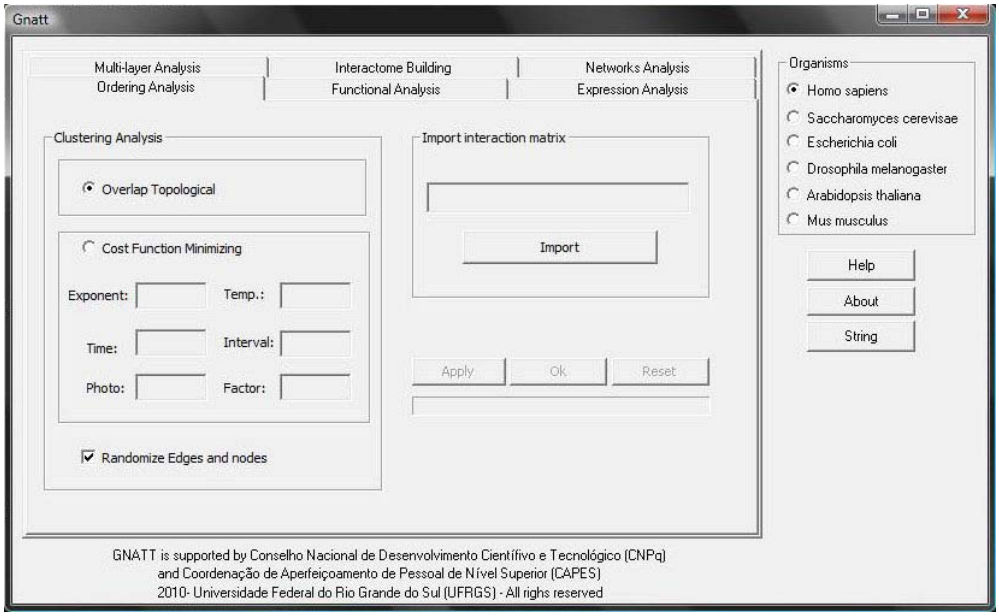


Fig. 5.1: *Interface de análise de clusterização do GNATT.*

função custo, na equação 5.1. Sendo assim, a equação tem a seguinte forma:

$$\begin{aligned}\mathcal{E} &= \sum_{i=1}^{N}\sum_{j=1}^{N}|i-j|^{\alpha}[|a_{i,j}-a_{i+1,j}|+|a_{i,j}-a_{i-1,j}| \\ &\quad +|a_{i,j}-a_{i,j+1}|+|a_{i,j}-a_{i,j-1}|]. \end{aligned} \tag{5.1}$$

- **Time:** é quantidade de passos de Monte Carlo do processo de reorganização.
- **Photo:** é intervalo, em número de passos de Monte Carlo, para salvar uma imagem da matriz de interação organizada.
- **Temp.:** é o valor da temperatura inicial para o processo de *annealing*.
- **Interval:** é o intervalo, em passos de Monte Carlo, entre os decréscimos de temperatura do processo de *annealing*.
- **Factor:** é o fator de decréscimo aplicado à temperatura em intervalos de passos de Monte Carlo, responsável por gradativamente reduzir a temperatura durante o processo de *annealing*.

A tabela 5.1 mostra o formato de entrada para análise de clusterização, onde a primeira linha tem o nome “Edges” seguida de linhas que formam duas colunas. Cada linha se refere

Edges	
Q0060	YMR197
YMR197	YRC167
YRC167	Q0045
Q0045	Q0060
Q0120	YLR197
⋮	⋮

Tab. 5.1: *Formato de entrada para a função de análise de clusterização.*

a uma interação, com as colunas listando os pares de proteínas que participam da interação. Antes da entrada destes dados, o usuário pode escolher o algoritmo de organização dos dados. Para ambas opções de organização, as saídas vêm nas formas apresentadas nas tabelas 5.2 e 5.3. Na primeira coluna são atribuídos números inteiros a cada gene da lista de entrada, contendo as interações. Observe que os genes rotulados por 1 e 2 da primeira coluna da tabela 5.2 são justamente os genes da primeira linha de interações da tabela 5.2.

Nodes(i)	Genes(i)
1	Q0060
2	YMR197
3	YRC167
4	Q0045
5	Q0120
⋮	⋮

Tab. 5.2: *Formato de um dos arquivos de saída: contém a enumeração inicial de todos os vértices da rede.*

Na tabela 5.3 temos o resultado do ordenamento. A primeira coluna contém a posição dos vértices na lista inicial, a segunda coluna informa a nova ordem dos vértices. Na terceira coluna estão os nomes dos genes ou proteínas, na quarta coluna, listamos a conectividade do vértice e a última coluna informa o coeficiente de clusterização. A próxima tabela 5.4 é uma saída do algoritmo de hierarquização, isto é um dendograma no formato do phylip. O **phylip**

Nodes(i)	Ordering(i)	Genes(i)	k(i)	c(i)
1	1	Q0060	1	0.00
3	2	YMR167	1	0.00
2	3	YRC197	2	0.017
127	4	YRC234	2	0.125
457	5	YMC164	3	1.000
⋮	⋮	⋮	⋮	⋮

Tab. 5.3: *Formato de saída contendo a reorganização da lista de gene. Assim, por exemplo, o gene YMR167, inicialmente na posição 3, depois de reordenado o genoma, encontra-se na posição 2.*

(*Phylogeny Inference Package*) é um pacote de 35 programas que auxiliam na inferência de árvores evolutivas [FEL 97]. O GNATT considera também o a matriz de *overlap* topológico,

⋮
(
150,
(
(
136,
(
127,
(
41,
54
)
⋮

Tab. 5.4: *Saída de dados no formato de phylip para as redes ordenadas pelo overlap topológico.*

que é uma matriz de números reais $N \times N$. O resultado é apresentado como na tabela 5.5.

A saída contém 6 colunas, sendo que as duas primeiras informam as posições na lista de entrada (Node(i) e Node(j)), seguidos de duas colunas que dão as posições dos genes na lista final, (ordering(i) e ordering(j)), a quinta coluna é o valor do *overlap* topológico $O(i, j)$ entre os pares de genes da linha, e a última coluna informa se existe (1) ou não (0) interação entre os genes.

Node(i)	Node(j)	Ordering(i)	Ordering(j)	top_ov(i,j)	link(i,j)
1	1	144	144	0.0000	0
1	2	144	141	0.9062	1
1	3	144	177	0.0714	0
1	4	144	7	0.0000	0
1	5	144	76	0.0000	0
1	6	144	35	0.0000	0
⋮	⋮	⋮	⋮	⋮	⋮

Tab. 5.5: *Resultado do algoritmo de overlap topológico.*

Nodes(i)	Nodes(j)	Nodes(j)	Nodes(i)
10	8	8	10
24	7	7	24
12	4	4	12
7	2	2	7
5	2	2	5
⋮	⋮	⋮	⋮

Tab. 5.6: *Este é o formato de saída com os dados para construir a matriz de interação após o ordenamento, com a informação de quais elementos da matriz assumem o valor 1. Por conveniência, a terceira e quarta colunas repetem a infomação da primeira e segunda colunas, refletindo o caráter simétrico da matriz de interações.*

5.2 Análise Funcional

Esta função tem como objetivo localizar no ordenamento genes relacionados com as diferentes processos metabólicos do organismo, figura 5.2, como termos do Gene Ontology. Considerando os termos do GO, temos três ontologias independentes: processo biológico, componente celular e função molecular. O GNATT possibilita a localização dos genes segundo o GO para os seguintes organismos: *Homo sapiens*, *Arabidopsis thaliana*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Mus musculus*, *Drosophila melanogaster*.

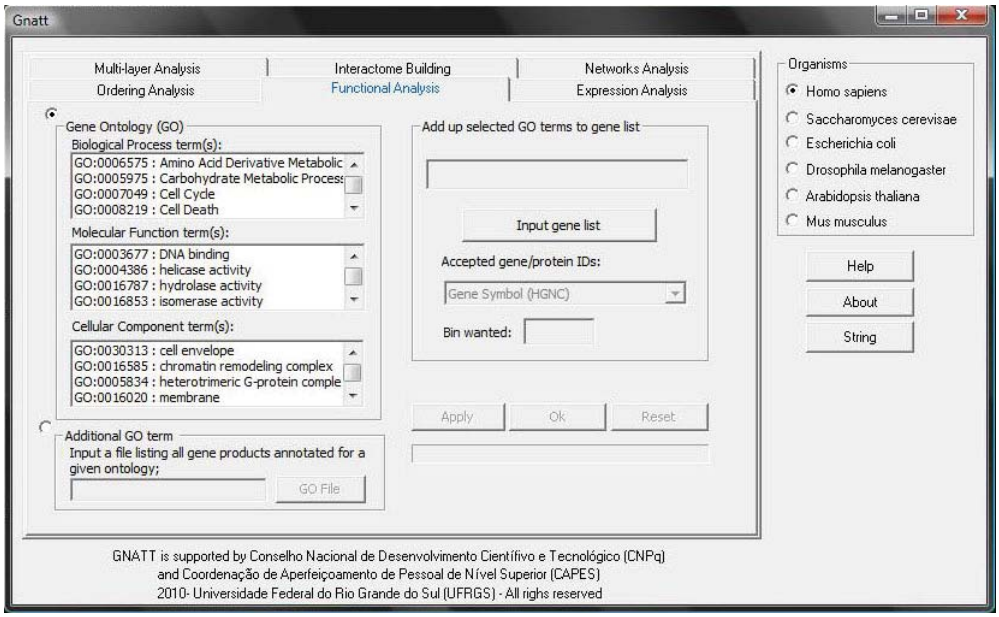


Fig. 5.2: *Interface de análise de funcional do GNATT.*

O GNATT, porém, abre a possibilidade de novas classificações. Para tanto o usuário deve utilizar a opção "Additional GO Term" e nesta função podem ser adicionados dados relativos tanto a novas classificações como a outras espécies: o *software* não fica preso somente nestas seis espécies. Em "Add up select GO terms to gene list" o usuário entra com os dados no formato da tabela 5.3 e em "Additional GO term" no formato da tabela 5.7. É ainda necessário entrar com o identificador dos genes em "Accepted gene/protein IDs", que é diferente para diferentes espécies. O tamanho da janela deve ser definido em "Bin Wanted" (w), lembrando que a largura da janela é $2 * w + 1$.

A Análise Funcional gera dois tipos de saídas, como nas tabelas 5.8 e 5.9 : O primeiro é

Genes
Q0060
YMR197
YRC167
Q0045
Q0120
⋮

Tab. 5.7: *Formato dos dados para a opção "Additional GO Term".*

igual ao formato da tabela 5.3 com adição das funções analisadas, onde existe "1" significa que tal gene existe na função em questão. No caso contrário é atribuído zero. No formato da tabela 5.9 a primeira coluna é a posição do gene/proteína no ordenamento e seu respectivo valor médio calculado sobre uma largura de janela $2 * w + 1$.

Nodes(i)	Ordering(i)	Genes(i)	k(i)	c(i)	Nome da Ontologia	⋯
10	1	Q0060	1	0.00	1	⋯
67	2	YMR197	1	0.00	1	⋯
37	3	YRC167	2	0.017	0	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Tab. 5.8: *Formato de saída das funções biológicas não mediados sobre os intervalos de janela.*

Ordering(i)	Nome da Ontologia	⋯
1	0.1493	⋯
2	0.1664	⋯
3	0.1843	⋯
4	0.2032	⋯
5	0.2230	⋯
⋮	⋮	⋱

Tab. 5.9: *Formato de saída das funções biológicas janeladas.*

5.3 Análise de Expressão

Esta função segue a mesma lógica da Análise Funcional, figura 5.3. A diferença é que se usa informação de microarranjos, portanto são dados referente aos níveis de expressão dos genes, obtidos em experimentos. Nesta função temos vários dicionários sonda-proteína/gene para plataformas AFFYMETRIX, que são largamente utilizadas para a construção de *chips* de microarranjos. O dicionário é necessário pois o ordenamento é realizado com uma rede gênica/protéica, portanto com nomes de genes/proteínas. O microarranjo é uma grade de sondas, cada sonda é capaz de se ligar a um gene/proteína. Os dicionários contidos neste software são para as seis espécies listadas, mas o usuário pode usar dados provenientes de microarranjos para outras espécies.

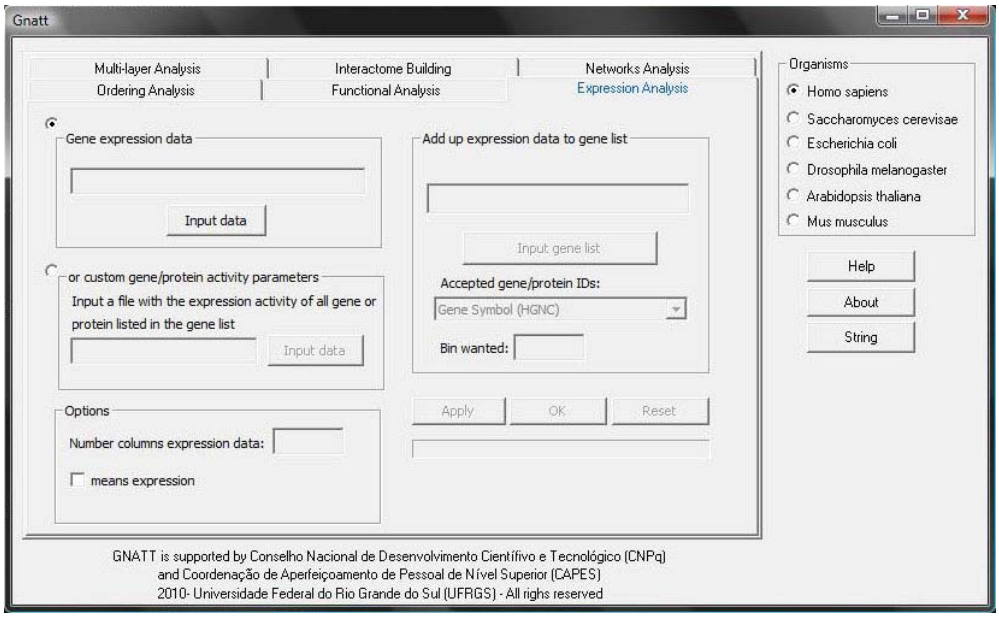


Fig. 5.3: *interface de análise de expressão do GNATT.*

O usuário insere seus dados de microarranjo no formato da tabela 5.10, cada linha corresponde a uma sonda e o seu respectivo valor de expressão.

No caso em que o usuário utilize outras espécies, é necessário que seja utilizado o nome dos genes/proteínas da rede no lugar do nome das sondas, conforme mostra a tabela 5.11.

O resultado da análise de expressão consiste num arquivo com formato da tabela 5.12: na primeira coluna temos o ordenamento dos genes, na segunda coluna, o número do gene antes

Probe(i)	Signal(i)
1007_s_at	1103.4
1053_at	20.4
117_at	12.7
121_at	237.4
1255_g_at	15.5
⋮	⋮

Tab. 5.10: *Formato de entrada para os dados de expressão gênica usando microarranjos da plataforma AFFYMETRIX.*

Gene(i)	Expression(i)
HMGB1	1601.42
NFKBIA	235.95
IKBKB	907.93
IKBKG	113.00
CHUK	4.20
⋮	⋮

Tab. 5.11: *Formato de entrada para os dados de expressão gênica usando microarranjos independente da plataforma.*

de ser ordenado, na terceira coluna encontramos o nome do gene do ordenamento, a quarta coluna corresponde ao valor do nível de expressão bruto do gene e, finalmente, a quinta coluna temos a média do valor de expressão gênica sobre a largura de janela selecionada pelo usuário.

5.4 Criação de Interatoma

O GNATT possui interatoma de seis espécies, figura 5.4, estes dados foram retirados do STRING. Para a construção dos interatomas estão disponíveis todos os métodos de predição de interação que são apresentados no site do STRING, como descrito na seção 1.5.1. O

Ordering(i)	Node(i)	Gene(i)	Expression(i)	Bin(i):10
1	150	HMGB1	1601.42	361.40
2	136	NFKBIA	235.95	368.55
3	127	IKBKB	907.93	333.81
4	41	IKBKG	113.00	352.95
5	54	CHUK	4.20	316.44
⋮	⋮	⋮	⋮	⋮

Tab. 5.12: *Formato de saída da análise de expressão.*

usuário seleciona os métodos de predição que ele deseja ("prediction methods") , e indica na caixa "confidence score" o escore desejado. O formato de saída da rede é idêntico ao formato da tabela 5.1, é possível alterar o formato de gravação da saída, no caso o usuário pode escolher se deseja que os dados sejam salvos nos formatos csv, tabulação, medusa e do pajek.

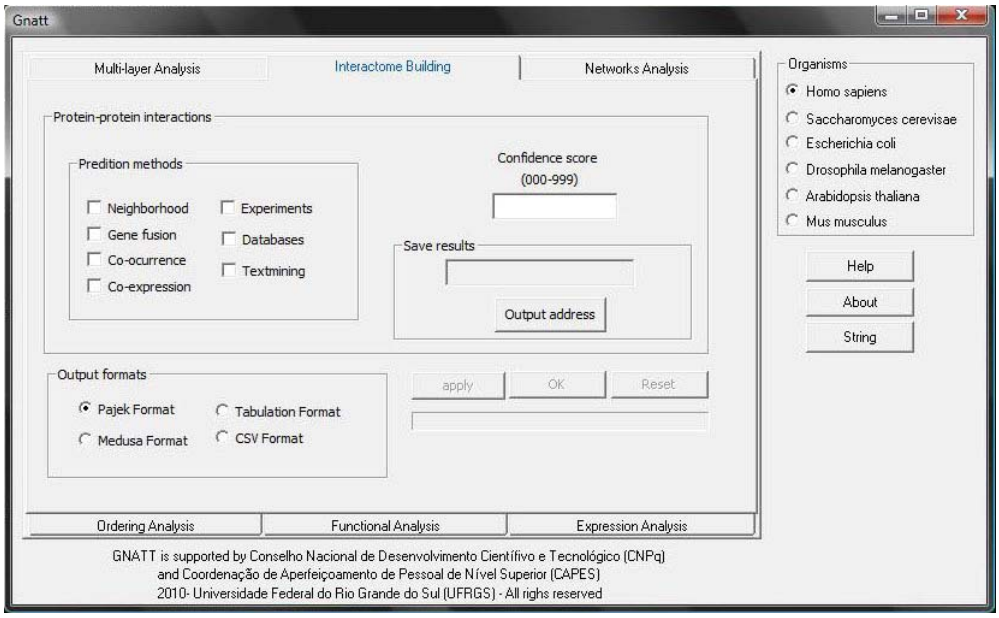


Fig. 5.4: *Interface de criação de interatoma do GNATT.*

5.5 Busca por Camadas

Nesta opção, figura 5.5, o usuário entra com duas listagens, uma é a matriz de associações (interações) e uma lista de proteínas/genes a serem pesquisados dentro da matriz de associação. Digamos que o entrada seja a matriz de interação do *Homo sapiens* e o usuário está interessado nos genes TP53, CFL1, e NFKB. Na lista de proteínas o usuário coloca o nome dos genes desejados e seleciona a vizinhança desejada. Se for escolhido "camada 1", a saída traz informação sobre a existência ou não das interações entre estes genes. Ao selecionar "camada 2" teremos os vizinhos dos três genes, no caso de "camada 3" teremos os primeiros e segundos vizinhos e assim por diante até "camada 5" . Esta ferramenta é usada para construir redes a partir de um ou mais vértices. O formato de entrada para os dados da rede para ser pesquisado é igual ao visto na tabela 5.1 e o formato do arquivo que contém os vértices a partir dos quais vamos construir a rede segue o modelo da tabela 5.7.

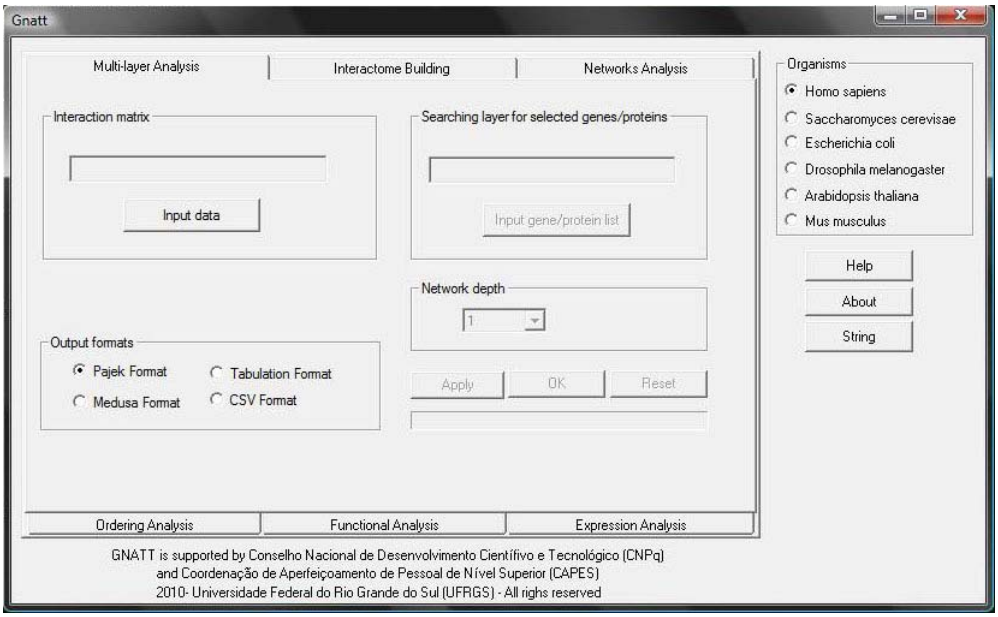


Fig. 5.5: *interface de análise de busca por camadas.*

5.6 Estatísticas de Rede e Modularidade

Nesta função, figura 5.6, é possível suavizar os dados de coeficiente de clusterização, conectividade, coeficiente de clusterização médio dos vizinhos e conectividade média dos vizinhos sobrepostos no ordenamento. O usuário define uma largura de janela (w) que será aplicada sobre os dados. Calcula-se a média dos dados dentro da janela e atribui-se o valor obtido ao vértice do centro da mesma (vide seção 2.4). Também é possível calcular a modularidade de janela (vide seção 2.3). Em "Bin Wanted" o usuário informa a largura da janela. O formato de entrada dos dados é idêntico ao visto na tabela 5.6. Também é possível calcular as estatísticas de $p(k)$, $C(k)$ e $K(k)$ da rede.

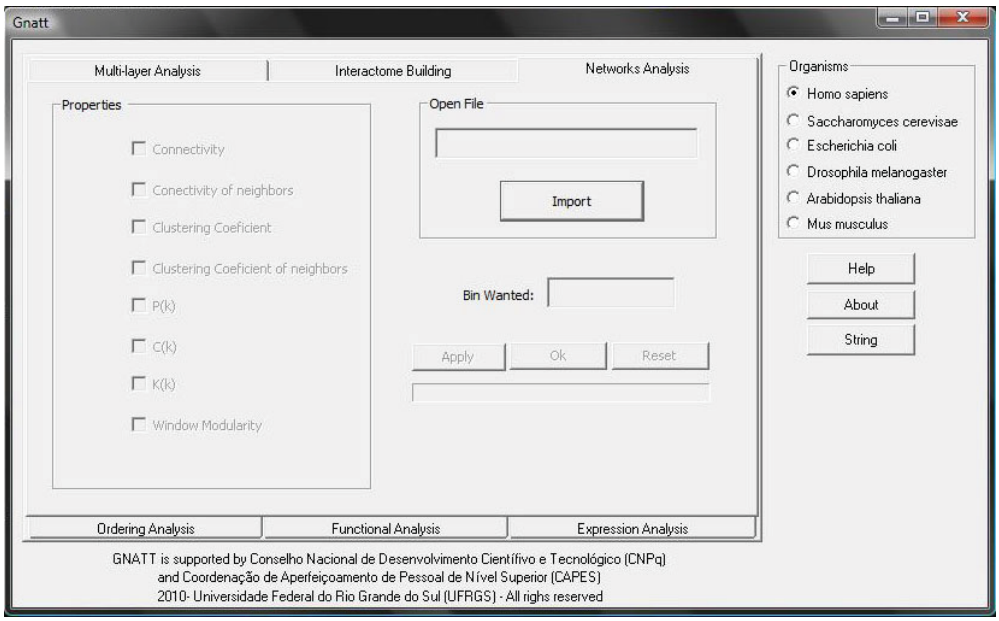


Fig. 5.6: *interface de cálculo de estatísticas de rede e modularidade.*

Capítulo 6

Conclusões

6.1 Conclusões

Neste trabalho foi apresentado um método de ordenamento de genes em uma lista usando a matriz de interação proteína-proteína. A figura 2.6 apresenta seis espécies cujo o genoma foi ordenado pelo método aqui proposto. A nossa metodologia de organização mostrou-se eficiente na agregação de genes de forma hierárquica. Com isso queremos dizer que a probabilidade de interação entre dois genes separados por n posições no ordenamento decai exponencialmente com n. A consequência é que podemos identificar genes mais próximos como pertencentes ao mesmo módulo além de reconhecer módulos mais próximos como pertencentes a um mesmo módulo de módulos. Outros algoritmos se propõem a encontrar clusters de genes, mas não os ordenam em uma lista.

Os dados referentes às interações proteína-proteína são frutos de vários anos de pesquisa, e o fato de organizá-las faz com que se revelem estruturas que antes não eram percebidas. O CFM lapidou estas redes de maneira que fossem reveladas estruturas modulares que não eram vistas pelos algoritmos para encontrar clusters em redes [dN 05, RA 08, AND 08, FRU 91, DEE 06, HAK 07, SAI 04, SPI 03, SPI 06, ENR 02, HOO 05, BAD 03].

Considere uma imagem bidimensional composta por um número muito grande de pixels binários (branco e preto). Tal imagem pode ser percebida por nossos olhos como apresentando todas as nuances de cinza. Isso acontece porque o olho humano percebe muitos pequenos pixels como um só, onde a intensidade de cinza depende da proporção entre o número de pixels brancos e pretos. De uma certa forma isso significa uma suavização do

perfil de brancos e pretos como se fossem feitas médias sobre uma vizinhança suficientemente grande de pixels para que muitos diferentes tons de cinza sejam possíveis, e suficientemente pequena para que o olho humano perceba a imagem como um contínuo.

No exemplo da imagem bidimensional, o pixel deve está correlacionado com a sua vizinhança para existir tal suavização. Analogamente, o CFM ordena uma lista de genes com o objetivo de encontrar a melhor correlação funcional de um gene com a sua respectiva vizinhança ao longo de todo o genoma, e assim produzindo estruturas modulares.

O uso de duas ferramentas *David tools* e *Gene Ontology* confirmaram que as estruturas modulares correspondem a módulos com significado biológico como apresentado na figura 2.12. Deve-se ter em mente que as interações já carregam essa informação biológica em suas conexões (arestas), e o CFM fez com que esta organização emergisse.

Sobrepondo os dados de transcrição (expressão gênica) sobre o ordenamento dos genes e suavizando tais dados com o uso de médias sobre vizinhanças de tamanho adequado, obtemos um perfil de transcrição - o transcriptograma. E o reconhecimento dos Processos Biológicos no ordenamento é de fundamental importância para interpretação do transcriptograma, já que alteração do perfil com respeito a um controle em um dado intervalo possibilita identificar os processo biológicos alterados.

Através do transcriptograma sabemos quais módulos ou genes estão alterados ou não nas células ou tecidos. Analisamos dois experimentos. Com base nas informações dos módulos foi possível obter as mesmas conclusões que os pesquisadores e, adicionalmente, foi possível extrair mais informação. Na referência [FRY 03], Fry *et al.* chegam à conclusão de que houve uma alteração de apenas 4% do genoma. Isto porque foram consideradas alteradas somente as expressões gênicas com intensidade duas ou mais vezes maior ou duas ou mais vezes menor do que as intensidades controle. No entanto, os genes têm sua dinâmica determinada por uma rede de outros genes. Assim, uma alteração moderada em intensidade de muitos genes pode causar efeitos não desprezíveis no metabolismo celular. O transcriptograma deste experimento mostra que houve uma alteração da expressão gênica do genoma inteiro.

O transcriptograma releva um perfil que considera todo e qualquer nível de expressão por menor que ele seja. Os módulos de genes medianamente superexpresso ou subexpressos parecem ter uma importância fundamental no metabolismo da célula e não devem ser

desconsiderados.

Durante a execução do presente projeto, bem como de outros correlatos, fomos encontrando tarefas que exigem a presença simultânea de físicos e bioquímicos. Tais tarefas podem se subdivididas em dois grupos. Aquelas onde os critérios, hipóteses ou resultados são discutidos, comparados e questionados, e um segundo grupo, que tratam da organização dos dados segundo alguma hipótese de trabalho. Para este segundo grupo, é possível operacionalizar as tarefas necessárias por meio de aplicativos computacionais que em muito facilitam sua realização. Para isso, este projeto contempla a criação e disponibilização em rede do GNATT.

O GNATT é uma ferramenta que está sendo desenvolvida para facilitar toda a análise que foi realizada ao longo deste trabalho. O software possuí os dois métodos de hierarquização de redes (CFM e *overlap* topológico), ele foi construído de maneira que qualquer usuário não possua nenhuma dificuldade em utiliza-lo.

6.2 Perspectivas

Até o presente momento, todo o estudo foi realizado usando o organismo *Saccharomyces cerevisiae*, que é uma levedura, um organismo unicelular. E os nossos resultados têm sido coerentes com as outras técnicas de análises utilizadas pelos autores citados no capítulo 4. Já estamos implementando a análise para *Homo sapiens*, que é um organismo totalmente diferente da *Saccharomyces cerevisiae*.

6.2.1 Melhor largura da janela

A largura da janela é uma questão importante, através dela o pesquisador decide se quer trabalhar com pequenos ou grandes módulos. Estamos trabalhando para informar qual é a melhor janela para o estudo de módulos de módulos. A Rugosidade nos diz qual a melhor largura para o estudo de módulos simples.

Varremos o ordenamento com todas as possíveis larguras de janelas, desde largura 1 até a largura da quantidade de nós da rede. Com estes dados, construímos mapas de cores que são matrizes quadradas (numero de proteínas x largura possíveis). Na figura 6.1 apresentamos

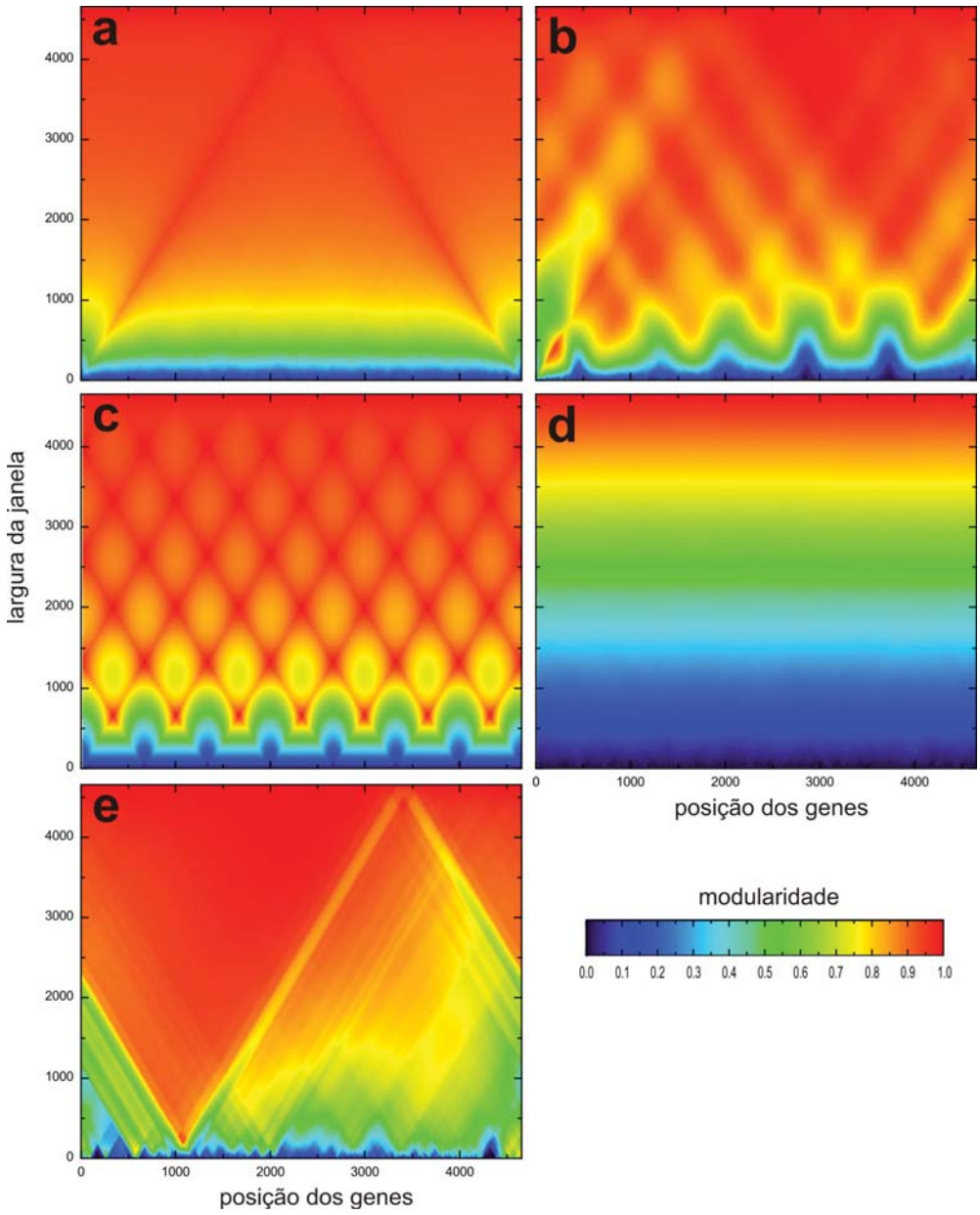


Fig. 6.1: *Mapas de modularidade para (a) rede exponencial, (b) rede exponencial modular (c) rede modular, (d) rede aleatória, (e) rede da Saccharomyces cerevisiae.*

seis matrizes, onde no eixo horizontal temos a posição relativa das proteínas e no eixo vertical a largura das janelas.

Para as redes **(a)** e **(d)** não existe nenhum módulo para qualquer largura de janela. A rede **(c)** apresenta picos azuis, verdes, verrmelhos que são módulos, quanto maior próximo do vermelho mais alto é o valor da modularidade (agrupamento de módulos). Em **(c)**, por

volta da largura $w = 665$ surgem os máximos (pontos vermelhos), em $w = 1330$ aparecem outros máximos, e a cada passo de 665 os picos são deslocados novamente, e temos um padrão de grade.

Para a levedura **(e)** existe uma banda de azul que são pequenos módulos (ver janela $w = 71$), a presença do primeiro pico vermelho aparece em torno de $w = 200$ e $w = 300$ que é o pico de tradução. Nota-se uma divisão da rede, a esquerda se forma um super módulo muito rapidamente, enquanto que na parte direita a evolução é mais lenta. Ainda estamos estudamos estes mapas para extrair a melhor janela.

6.2.2 Estudo de Câncer

Com a implementação de ordenamento para *Homo sapiens*, é possível construir transcriptogramas com dados oriundos de células cancerosas. Através desta metodologia agregada aos experimentos, poderemos ver qual módulo pode estar expressando mais ou menos em relação à célula normal. Também podemos buscar uma correlação em nível de expressão e estadiamento do câncer.

6.2.3 Ferramenta *on-line*

Existe uma carência de metodologias para análise de expressão gênica de genoma completo. O nosso objetivo é disponibilizar uma versão mais completa do GNATT associado ao VIACOMPLEX e GenPlast (em fase de desenvolvimento) em uma versão *on-line*. Essa ferramenta irá conter várias rotas metabólicas contruídas com STRING e KEGG, onde o usuário poderá montar as redes através da adição de rotas metabólicas normais ou alteradas (rotas de doenças) e ordená-las, logo após será possível criar transcriptogramas. Estes transcriptogramas poderão ser baixados e conterão uma análise estatística.

6.2.4 Ordenamentos bidimensional

O ordenamento apresentado neste trabalho é unidimensional, com isto só é possível correlacionar um vizinho à esquerda e outro à direita de um gene. Se expandirmos o ordenamento para duas dimensões, o número de vizinhos que podem ser correlacionados aumenta para

4. Isso pertimirá uma visualização melhor do que a unidimensional, em duas dimensões é possível gerar um gradiente de cores que facilitará à análise visual das regiões ativas e desativadas da rede.

Apêndice A

Artigo Towards a genome-wide transcriptogram: the *Saccharomyces cerevisiae* case

Nucleic Acids Research Advance Access published December 15, 2010

Nucleic Acids Research, 2010, 1–12
doi:10.1093/nar/gkq1269

Towards a genome-wide transcriptogram: the *Saccharomyces cerevisiae* case

José Luiz Rybarczyk-Filho[1], Mauro A. A. Castro[1,2], Rodrigo J. S. Dalmolin[3], José C. F. Moreira[3], Leonardo G. Brunnet[2] and Rita M. C. de Almeida[1,2,*]

[1]Instituto de Física, [2]National Institute of Science and Technology for Complex Systems and [3]Departamento de Bioquímica, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, 91051-970 C.P. 15051, Porto Alegre, Brazil

Received October 4, 2010; Revised November 16, 2010; Accepted November 22, 2010

ABSTRACT

Analysis of genome-wide expression data poses a challenge to extract relevant information. The usual approaches compare cellular expression levels relative to a pre-established control and genes are clustered based on the correlation of their expression levels. This implies that cluster definitions are dependent on the cellular metabolic state, eventually varying from one experiment to another. We present here a computational method that order genes on a line and clusters genes by the probability that their products interact. Protein–protein association information can be obtained from large data bases as STRING. The genome organization obtained this way is independent from specific experiments, and defines functional modules that are associated with gene ontology terms. The starting point is a gene list and a matrix specifying interactions. Considering the *Saccharomyces cerevisiae* genome, we projected on the ordering gene expression data, producing plots of transcription levels for two different experiments, whose data are available at Gene Expression Omnibus database. These plots discriminate metabolic cellular states, point to additional conclusions, and may be regarded as the first versions of 'transcriptograms'. This method is useful for extracting information from cell stimuli/responses experiments, and may be applied with diagnostic purposes to different organisms.

INTRODUCTION

Genome-wide expression data consist of expression levels of thousands of genes and the joint analysis of the whole data represents a challenge. The usual approaches compare expression levels of modified cellular stages relative to those of a pre-established control. The genes are then ranked by the variations in expression relative to the control and those genes that present the most significant alterations (highest or lowest) are chosen to be further analyzed. However, genes have their expression dynamics determined by a network of other genes and moderate alterations on many interacting genes may cause measurable effects on cell metabolism. These effects may be overlooked when using the maximally altered level criterion but, on the other hand, the great amount of data may prevent a more accurate analysis.

From a broader point of view, however, analysis of a great amount of information is not a novelty in scientific research. Even in everyday life events, people deal with amounts of data that largely exceeds their capacity to process. Indeed, data filtering to process only the most relevant information is an ability that saves time and energy and, probably, it has been repeatedly selected during evolution. A common example of data filtering can be given by a high-resolution photograph: although the digital file contains information on a huge number of pixels, much higher than the number of pixels in a computer screen, a picture of the whole object can still be produced on the screen. Image processing tools assign to each screen pixel some average of the information stored in a neighboring group of digital pixels, reducing the total information sent to the computer screen but still preserving global information. Observe that zooms may be applied to these pictures to obtain partial images such that, after a zoom, each screen pixel is assigned with the average of the information stored by a number of neighboring digital pixels. In other words, a huge collection of data relative to a whole phenomenon may be presented either by a coarser, global image or by a finer but partial image of the whole. In this example, the key point is the average of information stored on

*To whom correspondence should be addressed. Tel: +55 51 33086521; Fax: +55 51 33087286; Email: rita@if.ufrgs.br

neighboring pixels. Furthermore, the averaging over neighboring pixels also acts in the sense of neutralizing spurious fluctuations caused by some random external effect.

In this article we present a method to produce 'images' of gene expression data of whole genomes, by producing expression profiles for transcriptomes. The idea of the method is to consider averages of expression data over neighboring genes disposed on a line, as in the metaphoric example of the high-resolution photograph. In one hand, this procedure targets a global assessment of expression data of whole genomes. On the other hand, it requires the definition of gene neighborhood when disposed on a line, which is not straightforward.

Expression levels of different genes may differ by large amounts. Consequently a random list of genes generates plots of relative gene expression levels that fluctuate so wildly that very few, if some information, can be gathered from them. Techniques to extract information from wildly fluctuating general profiles consider averages taken over intervals of neighboring points. In the case where genes are ordered on a list following some criterion that favors clustering together interacting genes, the distance between any two genes on the list may correlate with the probability of mutual interaction, yielding then a natural criterion to define gene neighborhood on the list.

Many algorithms exist that find clusters of nodes in complex networks. These algorithms have been successfully applied to gene networks based on protein–protein interactions [see, for example, refs (1–4)]. However, they do not order genes on a list, but rather present the genes that belong to the same cluster in an arbitrary order. An exception is the clustering algorithm proposed by Barabási and collaborators (5,6), as we discuss in the following sections. Also, analysis of trancriptomes often cluster together genes by their co-expression, or co-variation in time, which implies that these cluster definitions depend on the stage the cell is going through or on the protocol used to produce the assessed sample.

Here we present a method for ordering a list of genes using the computational physics method known as Monte Carlo (7), that we call Cost Function Method (CFM). The aim is to cluster on a line interacting genes, such that the distance between two genes on the list correlates with the probability that they interact, that is, the probability that their protein products are associated in protein–protein association data bases as STRING (8). A first advantage is that the definition of these clusters is independent from the specific stage the cells are at a given moment, or the protocol they have suffered. The genome ordering we propose here defines a mathematical metric that correlates the distance between two genes on the list with their mutual influence. In this sense, the probability that two genes interact decreases with the distance between their localization on the ordered list, and an average of the expression levels over neighboring genes on this list dumps fluctuations and produces a smooth profile—that we call 'transcriptogram'. As we show in the following, the ordering is capable of clustering together genes belonging to terms of Gene Ontology: Biological Processes (9). Furthermore, expression profiles projected on the ordering give enough information on the global performance of a cell to discriminate different metabolic or biosynthetic processes, rendering a global assessment of cellular metabolism.

MATERIALS AND METHODS

We retrieved protein–protein interactions from STRING database (8th version) (http://string.embl.de/) (8), using 'experimental' and 'database' (95% of these interactions) added with 'neighbourhood', 'fusion', 'co-expression', and 'co-occurrence' evidences, String-score ≥ 0.800, comprising 4655 genes and 47 415 interactions.

Gene Ontology (GO) term enrichment was performed using DAVID bioinformatics resources (http://david.niaid.nih.gov) (10) to determine whether particular gene ontology terms occur more frequently than expected by chance in a given set of genes. We used default settings for the category GOTERM_BP_ALL, and selected those terms with $P < 0.05$ (for FDR no greater than 5%) representing central biochemical pathways/metabolic functions. From bit strings where the ith bit is set to 1(0) whenever the ith gene of an ordering is (not) listed in the GO term, we obtain smooth profiles by assigning to every gene the fraction of bits with value 1 in a window of size w, centred on the gene.

Yeast transcript expression data were obtained from YG_S98 array platforms (Affymetrix, Inc.), available at GEO database, Series GSE3431 (11) and GSE423 (12) (http://www.ncbi.nlm.nih.gov/projects/geo/). The transcriptograms are obtained by assigning to the ith gene the average of the expression values of its neighbors in a window of size w centred at the gene.

RESULTS

The starting point for the method is a randomly enumerated list of genes and the corresponding matrix specifying the interaction between the proteins. Here we consider gene or protein interaction as the physical and/or functional association presented by any pair of protein products. This body of information has been produced along the years by different researchers around the world and is magnificently organized and available at STRING database (8). We retrieved all protein–protein interactions described in that database inferred by 'experimental' and 'database' evidences for the organism *Saccharomyces cerevisiae*. Our final list comprises 4655 genes and 47 415 interactions.

For an ordered list with N genes, the interaction data may be organized in an $N \times N$ matrix M, where the matrix elements, $M_{i,j}$, are 1 or 0 depending on whether or not the ith and jth genes on the list interact. The result is a symmetric matrix of zeroes and ones with a null diagonal. We propose here an ordering algorithm that

favors the proximity of interacting genes by minimizing a cost function E assigned to each ordering, given as

$$E = \sum_{i=1}\sum_{j=1} d_{ij}\{\left| M_{i,j} - M_{i+1,j} \right| + \left| M_{i,j} - M_{i-1,j} \right| + \left| M_{i,j} - M_{i,j+1} \right| + \left| M_{i,j} - M_{i,j-1} \right|\}, \quad (1)$$

where, |..| stands for the positive value of the difference of the matrix elements located at neighboring sites and d_{ij} is proportional to the distance from the point (i,j) to the diagonal, that is, $d_{ij} = |i-j|$. This cost function increases with the number of interfaces between one and zero elements on the matrix and increases further when these interfaces are far from the diagonal. We remember that points (i,j) far from the diagonal present very different values for i and j, implying then interactions between distant genes on the ordering.

After starting with a randomly ordered gene list and its corresponding interaction matrix, the algorithm proceeds by randomly choosing a pair of genes and swapping their positions on the ordering. A new interaction matrix is produced for this new ordering and its cost is recalculated using Equation (1). If the cost decreases, the change is accepted. If the cost is increased by ΔE, the change is accepted with probability exp [$-\Delta$E /T], where T is a virtual temperature. We started with $T = 6 \times 10^5$ and every 100 Monte Carlo Steps (MCS) the temperature is lowered to 20% of its previous value. A MCS is a number of random choices equal to the number of elements in the system. This procedure is known as a 'simulated annealing' (13), and is intended to escape from metastable states. When changes are not accepted, they are discarded and a new gene pair is chosen to repeat the process. This procedure is repeated until the calculated value of cost is stabilized. See Supplementary Figure S1 for the plot of the cost function versus number of changes.

Figure 1 presents the interaction matrices relative to *S. cerevisiae* for the initial random gene ordering (Figure 1a), after ordering following the Dendogram clustering algorithm as proposed by Barabási and collaborators (5) (Figure 1b), and following the algorithm described above (Figure 1c). For each figure, vertical and horizontal axes give the relative gene positions on the ordering. See also Supplementary Data for the details for the Dendogram ordering. These positions are normalized, such that the ith gene on the list is assigned the position $\frac{i}{4655}$ on both vertical and horizontal axes. In these figures a black dot located at (i,j) indicates an association between the gene in position i on the horizontal axis with the gene on position j on the vertical axis such that $M_{i,j} = 1$. All three configurations present the same number of black dots and represent the same information on protein–-protein association.

The difference between the figures stems in the different localizations of the genes on the axes. The randomly ordered gene list distributes uniformly the interaction-representing-dots over the whole matrix surface. After Dendogram ordering, some black dots are concentrated on the main diagonal with some large clusters, while after CFM ordering the black dots concentrate even nearer the diagonal, leaving the top left and bottom right corners free of black dots. These two corners represent interactions between genes located far apart on the list, since they represent matrix elements M_{ij} for which i and j are very different. Furthermore, the black dot clusters far from the diagonal, which are present in the interaction matrix representing the Dendogram ordering, indicate that there are many interacting genes belonging to clusters located far apart on that ordering.

A way of quantitatively characterizing the orderings is using the interaction probability $\rho(n)$ for two genes that are separated by n positions on an ordering. This probability is given by the relative number of black dots on diagonals distant by n pixels from the main diagonal on the interaction matrix and may be calculated as

$$\rho(n) = \frac{1}{N-n}\sum_{i=1}^{N-n} M_{i,\,i+n}. \quad (2)$$

Figure 2 presents $\rho(n)$ versus n for the Dendogram and CFM algorithms in log-linear plots. Observe that the Dendogram algorithm stabilizes $\rho(n)$ at a finite value as n increases, but the CFM algorithm yields an exponential decay (represented by a straight line in a log linear plot) for this probability. It implies that the probability that two genes interact decays exponentially with the distance between their locations on the CFM ordering, while stabilizing at a finite value ($\sim 10^{-3}$) in the case of the Dendogram ordering. For very short ranges, however, Figure 2b shows that the Dendogram algorithm concentrates more on the interacting genes, up to 20 genes distant; between 20 and approximately 600 genes apart the CFM concentrates more, between 600 an 1000 they present roughly the same interaction probability and, after that, the CFM ordering presents exponentially decreasing $\rho(n)$. We interpret this exponential decay in $\rho(n)$ for the CFM ordering as a correlation between interaction and localization of the genes on the ordering. This correlation yields to adequate averages over neighboring genes, allowing the smoothing out of wild fluctuations in the diverse profiles. For comparison we considered four artificially constructed networks whose results are presented on Supplementary Materials Online.

Gene ordering on a line is a frustrated process, in the sense that conflicts appear on how to order genes. It may happen that a gene interacts with two different clusters, say cluster A and B. This gene could be located near any one of the clusters or in some place in between. A criterion must be provided to resolve these conflicts. When favoring putting this gene together with, for example, cluster A, the blocks near the diagonal are more compact, but the price for that is the appearance of dots far from the diagonal, representing the interactions of the gene with cluster B. On the other hand, when the ordering method favors putting the gene in some place between clusters A and B on the ordering, the locations far from the main diagonal on the interaction matrix are free from black dots (there is no interaction M_{ij} such that i and j are very different) but the blocks near the diagonal are less compact.

While ordering, CFM algorithm acts to reduce a cost function by penalizing configurations with interactions

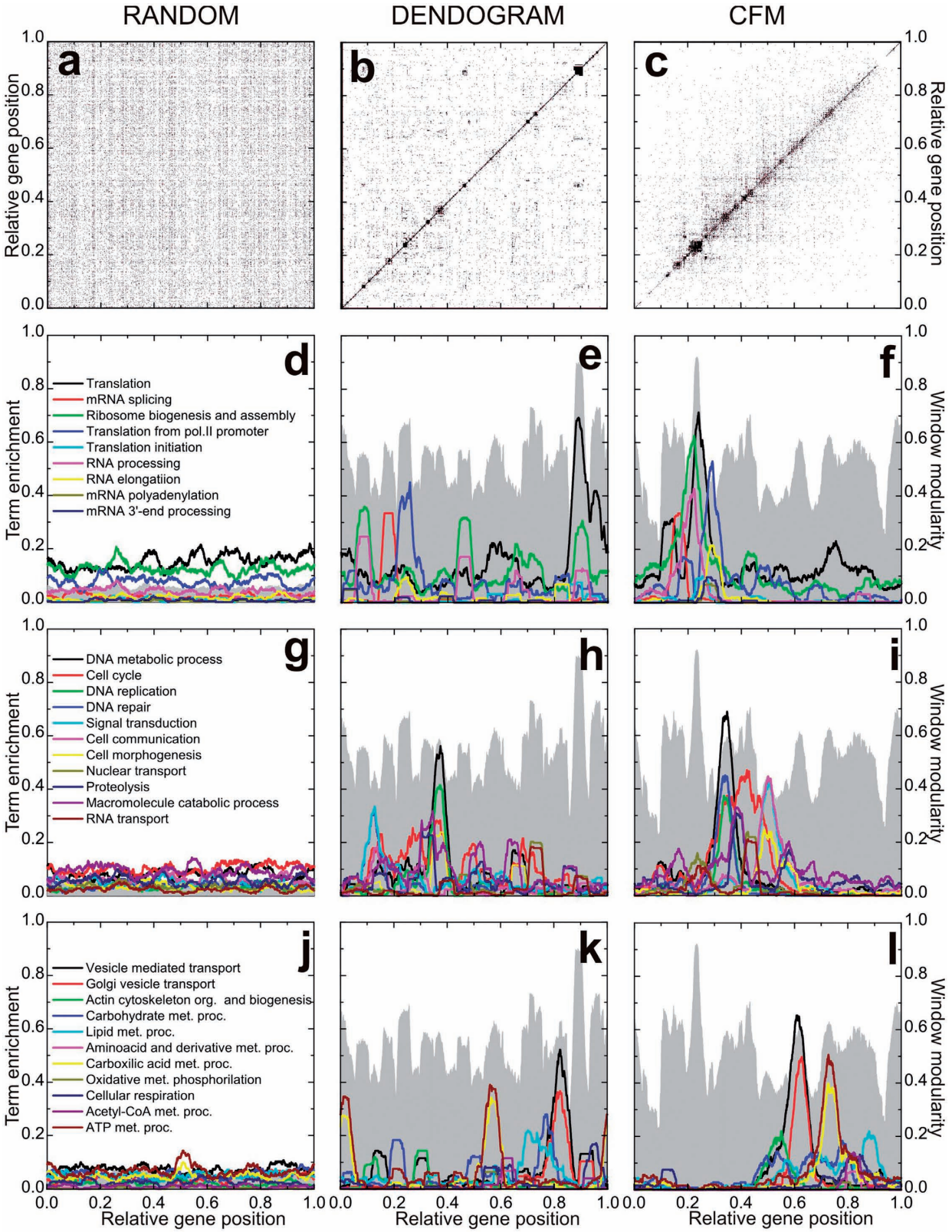


Figure 1. Protein–protein interaction matrix analysis algorithms. The axes relative to gene position have been divided by the total number of genes: 4655. (**a**) Random ordering. (**b**) Dendogram ordering algorithm. (**c**) Cost Function Minimizing (CFM) algorithm. (**d–l**) Projection of diverse terms of the Gene Ontology: Biological Process, as indicated in the right hand frame of each row. Gray landscape backgrounds: window modularity for the orderings. The maxima at the window modularity plots correspond to larger concentrations of black dots on the matrix representation, that is, intra-module interactions are more intense in these regions.

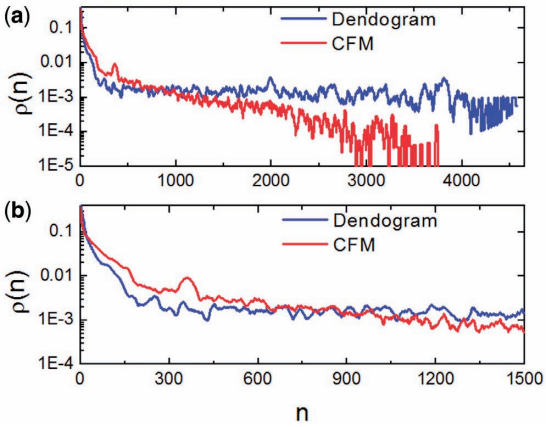


Figure 2. Interaction probability $\rho(n)$ as a function of n. This gives information on the quantity of links between genes as a function of their distance on the ordering. (**a**) On a log-linear plot and the whole interval, to evince the exponential decay of dot density on the CFM plots and (**b**) for a smaller interval in n to evince the behaviour near the main diagonal.

between genes located far apart in the ordering. This is done by the factor in Equation (1) that depends on the distance of a black dot (representing interaction between two genes) from the main diagonal. The consequence is reflected in the probability that two genes located n positions apart on the ordering, $\rho(n)$, decays exponentially with n.

The Dendogram method, on its turn, does not penalize strongly enough interactions between genes far apart on the ordering, favoring compact blocks near the main diagonal. The consequence is an interaction probability between genes, $\rho(n)$, that is large for small n, decreases fast for intermediate distances and then stabilizes at a constant value, as shown in Figure 2.

Window modularity

To further characterize the orderings, we have considered the window modularity for each gene on an ordering, defined as follows. For each gene on the ordering consider its $w/2$ neighbors to the left and its $w/2$ neighbors to the right, comprehending an interval of $w+1$ genes. The window modularity $W_w(i)$ for a gene, located at the ith position of the ordering, is defined as the ratio between the number of interactions that link any two genes in the interval (window) of size $w+1$ list, centered at the $i$thgene, and the number of interactions involving at least one gene in that window (14). That is,

$$W_w(i) = \frac{1}{\sum_{j=1}^{N} M_{i,j}} \sum_{j=\mathrm{mod}(i-\frac{w}{2},\,N)}^{\mathrm{mod}(i+\frac{w}{2},\,N)} M_{i,j}, \qquad (3)$$

where,

$$\mathrm{mod}(i+n,\,N) = \begin{cases} i+n & \text{if } i+n \le N \\ i+n-N & \text{if } i+n > N \end{cases} \qquad (4)$$

accounts for periodic boundary conditions to deal with genes near the ends of the list.

Window modularity strongly depends on the window size w. For example, for a window containing all genes of an ordered list, window modularity is one for every gene. However, when a gene is at the center of an interval that describes a highly interconnected cluster, its modularity decreases for windows smaller than the cluster size. This happens due to interactions connecting genes inside the window with genes outside the window but still belonging to the cluster. Also, genes that link different clusters present low modularity. On Figure 1d–l window modularity is represented as gray landscapes. There we have chosen $w = 251$. Plots for other values of window size are presented in Supplementary Data, as well as for other artificial networks for didactic purposes. The choice of the window size w depends on the desired accuracy for the peaks. In fact, as window size increases, the rugosity of the window modularity profile varies. It first increases, passes to a maximum and then decreases as w increases. Rugosity of a profile is defined as the standard deviation of the profile height and gives a measure of the amount of peaks and valleys. (See Supplementary Figure S4c and S4d). Here we choose $w = 251$ to have a more global description of the GO: BP terms. However, smaller windows may enhance accuracy for the modularity profile, as well as for expression data analysis (See Supplementary Figure S13 and S14 and discussion below).

Observe that window modularity in both Dendogram and CFM orderings present well defined peaks and valleys, indicating interacting modules. The random list presents a very low modularity for all genes. Taking random fluctuations as a null hypothesis, and estimating the standard deviations of random fluctuations from the random ordering modularity ($\sigma \sim 0.00735$), the probability that both CFM and Dendogram window modularity peaks and valleys are random is virtually zero, that is, peaks and valleys of heights of order 0.5 are more distant from the random average than 50 standard deviations of the window modularity distribution in a random ordering. The magnitude of both average and standard deviations for the window modularity may be directly estimated from the figures.

Although, the peaks in CFM and Dendogram orderings are similar in height, in the CFM ordering the valleys are deeper and the number of peaks separated by deep valleys is smaller. In fact, since there are valleys with different depth in the CFM ordering the peaks may be hierarchically defined: smaller clusters composing larger clusters.

Biological characterization

To assess the biochemical meaning of the orderings we have projected on the ordering information regarding the Biological Process terms from the Gene Ontology (GO) Database (9). We used 'DAVID' Bioinformatics Resources (10), as described in Materials and Methods, to obtain the GO terms of Biological Process Ontology that best represent each window modularity peak. After obtaining the representing terms, we calculated for each one a profile over the whole ordering. These GO term profiles are smooth functions of gene localization and give the fraction of genes that belong to the GO term in

windows of 251 sites around a given gene. See Figure 1d–l. For the randomly ordered list, no peaks are seen and no information can be gathered from these plots. For the ordering obtained using Dendogram algorithm, some peaks appear, but the ontology terms are not as concentrated as for the CFM algorithm. Again, having a null hypothesis of random fluctuations whose standard deviation is estimated from the random ordering projections, the probability that the peak values of the terms profiles presented by both CFM and Dendogram algorithm are random fluctuation is virtually zero, since they lay more distant from the random average than tens of standard deviations. See Supplementary Figures S8–S10 . Also, the CFM ordering successively locates classes of GO terms in an order that reproduces cell cycle: from right to left we first find terms associated with energy metabolism, followed by cell morphogenesis and cell communication, then GO terms related to vesicle transport and Golgi vesicle transport, then DNA replication and repair, and finally GO terms associated with RNA production and translation.

Network properties

The orderings may be characterized using the connectivity $k(i)$ and the clustering coefficient $c(i)$ of the ith gene on the ordering (15). The interaction matrix gives information on which pairs of genes interact. The connectivity $k(i)$ of the ith gene on the ordering is defined as the number of genes with which it interacts. On its turn, the clustering coefficient $c(i)$ is defined as the fraction of existing links between any two of the $k(i)$ neighbors of the ith gene, relative to the maximum possible number $k(i)[k(i)\text{-}1]/2$ of such links. Figure 3a and 3b presents the connectivity and clustering coefficient profiles for, respectively, the CFM and Dendogram orderings, obtained by taking the average of these quantities over windows of 251 sites. The connectivity profile of the CFM ordering shows that (i) genes with higher connectivity are more concentrated than the Dendogram ordering, presenting a high peak around the window modularity maximum at the region located at 0.2–0.3 on the horizontal axis. This region of the CFM ordering is rich with genes belonging to GO terms associated with translation, while the poorly connected genes are found at the ordering extremities. Also Figure 3 shows that (ii) the clustering coefficient decreases to very small values at the ordering extremities for the CFM ordering.

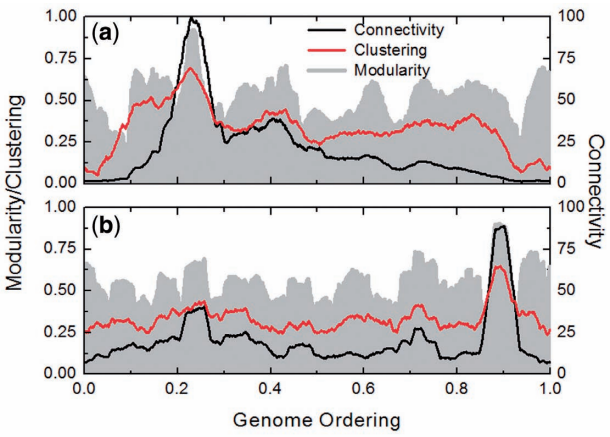


Figure 3. Connectivity and clustering coefficient for (**a**) CFM and (**b**) Dendogram ordering. The gray landscapes are relative to the window modularity. The window size is 251.

From now on we concentrate in analyzing the results for the CFM ordering. We sliced the CFM ordering in seven pieces, using the window modularity peaks as a guide (Figure 4e). The genes of each piece, together with the information on the interaction between these genes, are fed to Medusa application (16) and partial network graphs were produced, shown in Figure 4. The biological functions are mapped with GO terms. Observe that in this figure we are able to discriminate gene networks of related functions.

For example, networks p1, p2 and p3 (Figure 4a–c) are all associated with transcription and translation processes, as rRNA/mRNA processing and ribosome biogenesis and assembly. Network p4, also overlaps these functions (Figure 4d), represented by DNA repair/replication and cell-cycle regulation. All these four gene networks have in common the synthesis of biological polymers. By contrast, network p5 seems to be a single cluster, shifting the ordering to other biochemical classes (Figure 4f), such as cell communication and morphogenesis. The last two gene networks (Figure 4g–h) present a variety of functions, from actin cytoskeleton organization and vesicle transport to carbohydrate, lipid and amino acid metabolic processes.

A feature of the right side of CFM ordering is the presence of several intermediate products and ATP-producing pathways (e.g. carboxylic acid cycle and cellular respiration). The network structure is enriched with highly interconnected anabolic and catabolic pathways, which is consistent with the basic strategy of central metabolism to form ATP, electron carriers and precursors for the biosynthesis of more-complex molecules. Therefore, gene networks p6 and p7 are related to the production of both energy and the building blocks from which other biomolecules are made.

At the other end of the CFM ordering (the left side), the functional boundaries of the network structure seems to be better discriminated. There are sub-clusters associated with several processing steps that control the flow of genetic information in cells.

In summary, the metabolic pattern as organized by the CFM algorithm gives rise to a sound biochemical and functional ordering, where the closest gene networks are more interrelated than the distant ones.

The transcriptogram: projection of gene expression data

Now we analyze gene expression data for the yeast genome. We focus on experimental data available at Gene Expression Omnibus database, regarding microarrays presenting probes for almost all genome components. We have then projected the expression on the CFM ordering, always considering window averages,

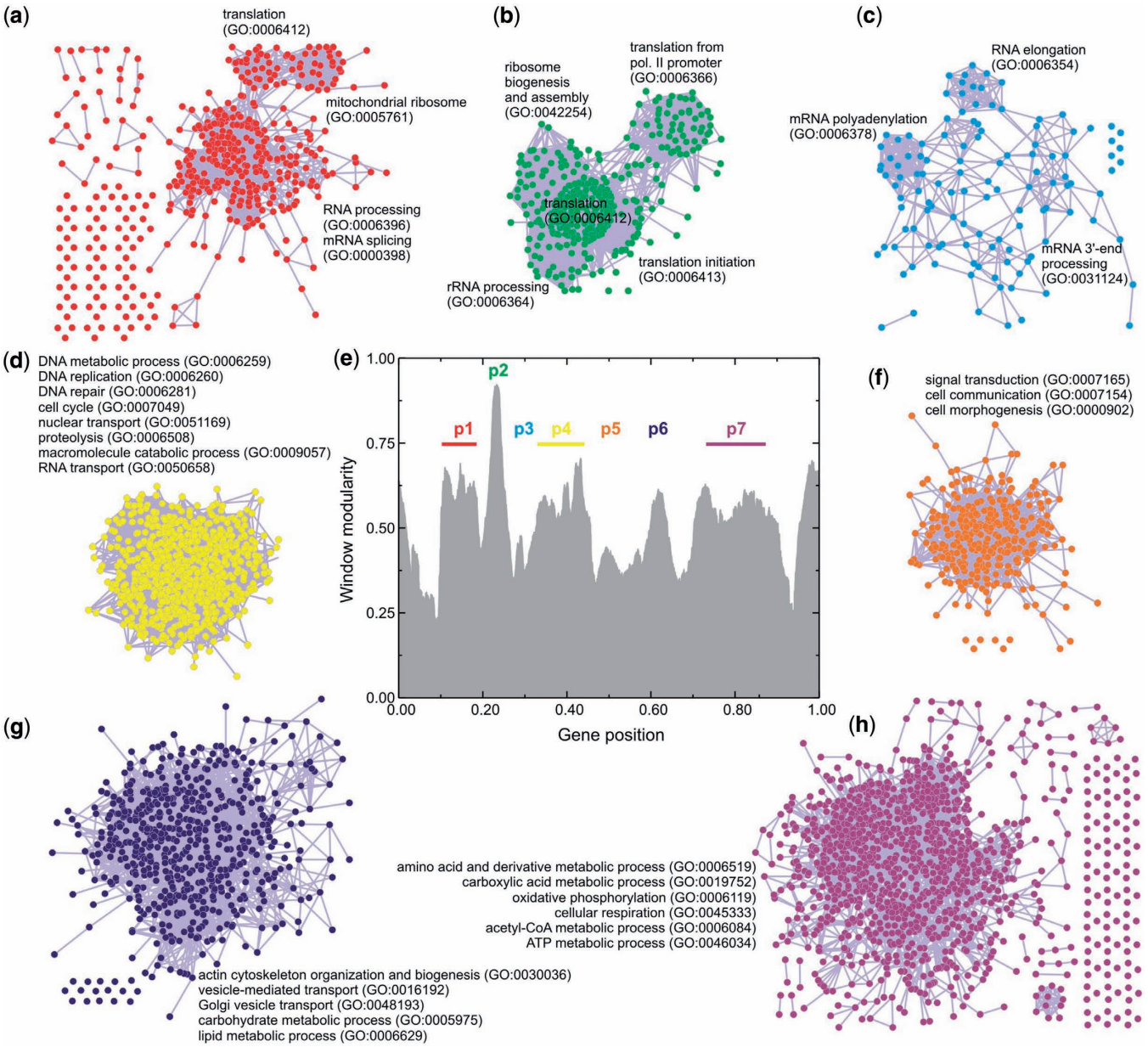


Figure 4. Graph representation of the CFM ordering. The axes relative to gene position have been divided by the total number of genes: 4655. (**e**) CFM ordering was sliced in seven pieces, using the window modularity peaks as a guide for this division. The genes of each piece, together with the information on the interaction between these genes, were fed to Medusa application to produce the network graphs. (**a**–**c**,**e**–**h**) Network graphs associated with each peak, whose biological functions are mapped with GO terms using 'DAVID' bioinformatics resources.

obtaining expression profiles that we call transcriptograms. Here we present transcriptograms for *S. cerevisiae* using data obtained from two different experiments.

The first one, as explained in a very nice paper by Tu *et al.* (11), considers expression data obtained from yeast continuous culture, in controlled conditions, where the concentration levels of dissolved O_2 are constantly monitored. These levels vary periodically in time and the transcription levels were measured for 12 different stages in three different dissolved O_2 concentration oscillation periods, summing up 36 transcription profiles.

Figure 5 presents the results concerning transcriptograms obtained using the CFM ordering. A movie presenting all 36 snapshots is available at Supplementary Materials Online, as well as the results for the Dendogram ordering. Figure 5a presents 21 transcriptograms (7 per cycle), taken at the instants represented by the colored (orange, blue and purple) dots on the plot of dissolved oxygen versus time in log-linear plot (Figure 5b). Each color is associated with one cycle. Figure 5a also presents the window modularity as a landscape, to guide the eye, and the distribution of three gene clusters as defined in Tu *et al.* paper based on sentinels genes: Ox (oxidative), R/B (reductive, building) and R/C (reductive, charging). Figure 5c–i present the relative expression profiles at different instants. The relative profiles were calculated taking as reference the average of the expression intensity for each gene

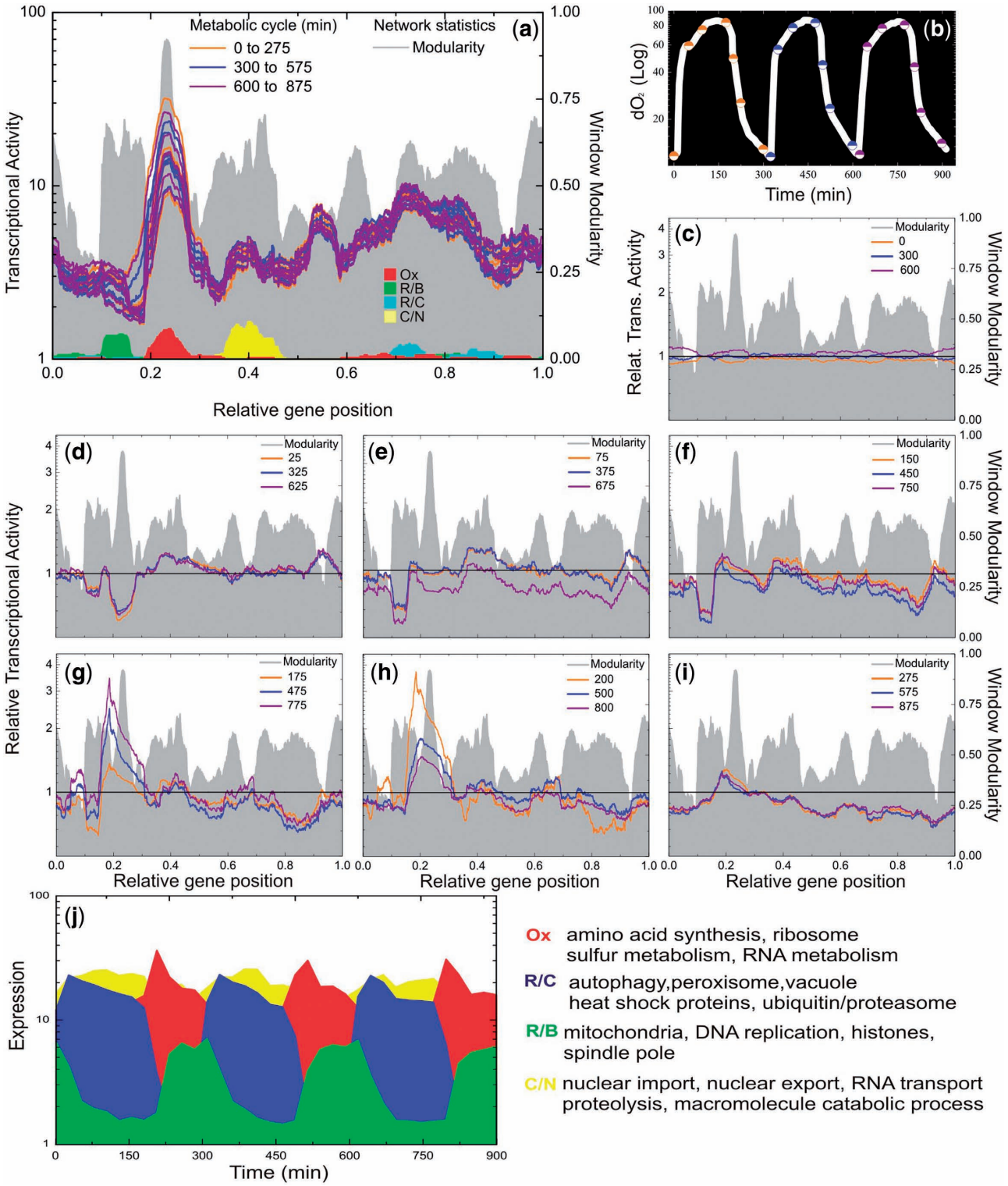


Figure 5. *Saccharomyces cerevisiae* transcriptograms. The axes relative to gene position have been divided by the total number of genes: 4655. (**a**) Microarray data available at Gene Expression Omnibus database were projected on CFM ordering to obtain the expression profiles, or transcriptograms. Each color is associated with one cycle, as shown in (**b**). Projections on the ordering were performed always considering window averages. To guide the eye, the window modularity is depicted as a landscape, together with the distribution of three gene clusters, as described previously, based on sentinels genes: Ox, oxidative; R/B, reductive, building; R/C, reductive, charging. Also, the distribution of the 40 genes whose expressions are maximally altered in the interval 0.35–0.45 of the CFM ordering. As discussed in the text, these genes are mainly related to catabolism of macromolecules and nuclear transport. (b) Plot of dissolved Oxygen versus time in log linear. Transcriptograms (seven per cycle), were taken at the instants represented by the colored (orange, blue and purple) dots. (**c–i**) Relative expression profiles. Transcriptograms were divided by the average expression values of the first state of the cycles (Time = 0, 300 and 600 min). c: represents the relative expression profile corresponding to the first dot of each cycle; d: represents the second dot of each cycle and so on. (**j**) Oscillations in expression levels of the sentinels genes: Ox, oxidative; R/B, reductive, building; R/C, reductive, charging, together with the most altered genes for the interval 0.35–0.45 of CFM ordering (yellow). These are average levels of the 40 most altered genes in each case.

presented at times equal to 0, 300 and 600 min, which represent the first stages of each cycle. We have divided each gene expression intensity for its respective average, and projected over the ordering after performing a 251-window average, as done for other quantities.

The expression profiles show different behaviours for the left and right hand side portions: the expression profile of left side peaks extremely abruptly at the intense burst of oxygen consumption, while the right side gradually rises when cells begin to cease oxygen consumption. According to the gene networks mapped in Figures 1 and 4, the left side embraces several energy-demanding processes, essentially represented by the synthesis of biological polymers. It requires abundant amounts of adenosine triphosphate (ATP), which is available in profusion at the respiratory phase. This interplay of metabolic pathways for energy production is compatible with the time ordering through the phases Ox, R/B and R/C as described in the original article (11).

Our results support the conclusion drawn by the authors based on the expression of 40 genes for each cluster, a small gene fraction available in yeast transcriptomes. Here, by the use of transcriptograms, we present the dynamic changes during the metabolic cycle assessing the complete information.

Moreover, the transcriptograms allow going further. There are more regions in the ordering that are significantly varying during the yeast life cycle. Figure 6 presents the transcriptograms together with the significance intervals for each point, given as the colored irregular bands. These significance bands have been calculated as follows. Taking the points at $T = 0$, 300 and 600 min as the reference (the initial stage of each respiratory cycle), we estimated the variation from the standard deviation of relative expression levels for each gene. The yellow band stands for relative gene expression levels that deviate from the initial stage average from 0 to 2 SDs. The pale pink band stands for regions where the relative expression levels deviates from the initial values from 2 to 4 SDs and the gray regions stand for deviations larger than 4 SDs. Because the expression levels of each gene present different values of standard deviations, these bands present irregular interfaces. Besides the regions pointed by Tu *et al.* the transcriptograms point to the interval from 0.35 to 0.45 as significantly varying during the respiration cycle (several standard deviations). See Figure 6 for times from 25 to 125 min. For illustration, we present in Figure 5j the average expression levels of the 40 genes that present the highest variations in this interval,

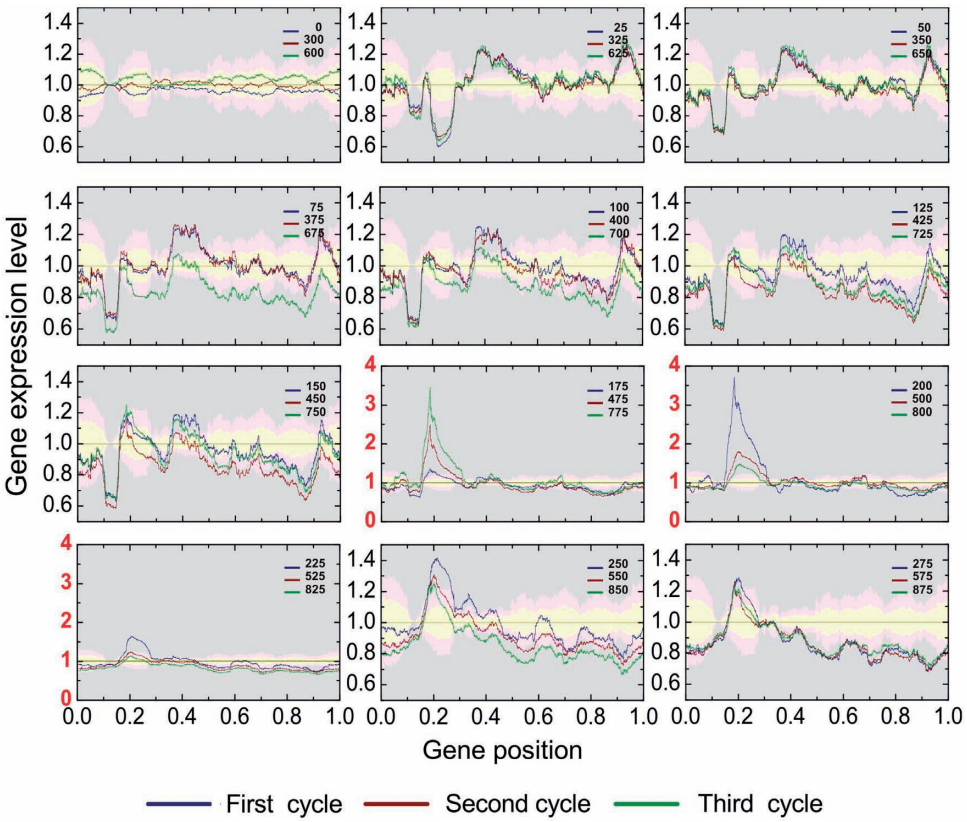


Figure 6. The respiration cycle transcriptograms and an estimate of the confidence intervals. The reference averages and their variations are estimated from the average and standard deviation for the relative expression level of each gene i in the initial state of the three respiration cycles. The yellow bands represent values that deviates from the average from 0 to 2 σ_i; pink bands stand for deviations from 2 σ_i to 4 σ_i, and the gray area stands for deviations larger than 4 σ_i. Each panel presents data relative to corresponding states of the three cycles. The ordering region in the interval 0.35–0.45 presents variations that are clearly in the gray region, mainly in times corresponding to $T = 150$, 450 and 4750 min, and $T = 175$, 475 and 775 min, which correspond to the final fermentation phase and the beginning of high consumption of O_2.

together with the average expression levels of the three groups of 40 genes presented by Tu *et al*. Although the oscillations in the expression levels are less intense than those found by Tu *et al*., they are still largely significant.

The density profiles for these 40 genes are represented in Figure 5a by the yellow peak. This is a group rich in genes belonging to macromolecule catabolic process terms or nuclear transport. In fact, these 40 genes belong to two different sub-peaks of peak 4 in Figure 6, that are made visible when we use a smaller window ($w = 101$ instead of $w = 251$), as shown in Supplementary Figures S13 and S14. The complete list of genes in the interval between 0.35 and 0.45 of the ordering may be found in Supplementary Table S1 and Supplementary Table S2 indicates the 40 genes that are maximally expressed in this interval.

The second experiment is the one by Fry, Sambandan and Rha (12), where the authors compare the transcription levels of *S. cerevisiae* wild-type with those of sgs1 mutants, when the samples are submitted or not to stress represented by the direct addition of 0.1% methyl methanesulfonate (MMS) and the cultures were incubated at 30°C for 1 h. Their conclusions from the results are that (i) under normal conditions the mutant present 4% of the genes with transcriptional levels altered by 2-fold or more and (ii) under the stressed conditions there is not any difference between the different lineages. Figure 7

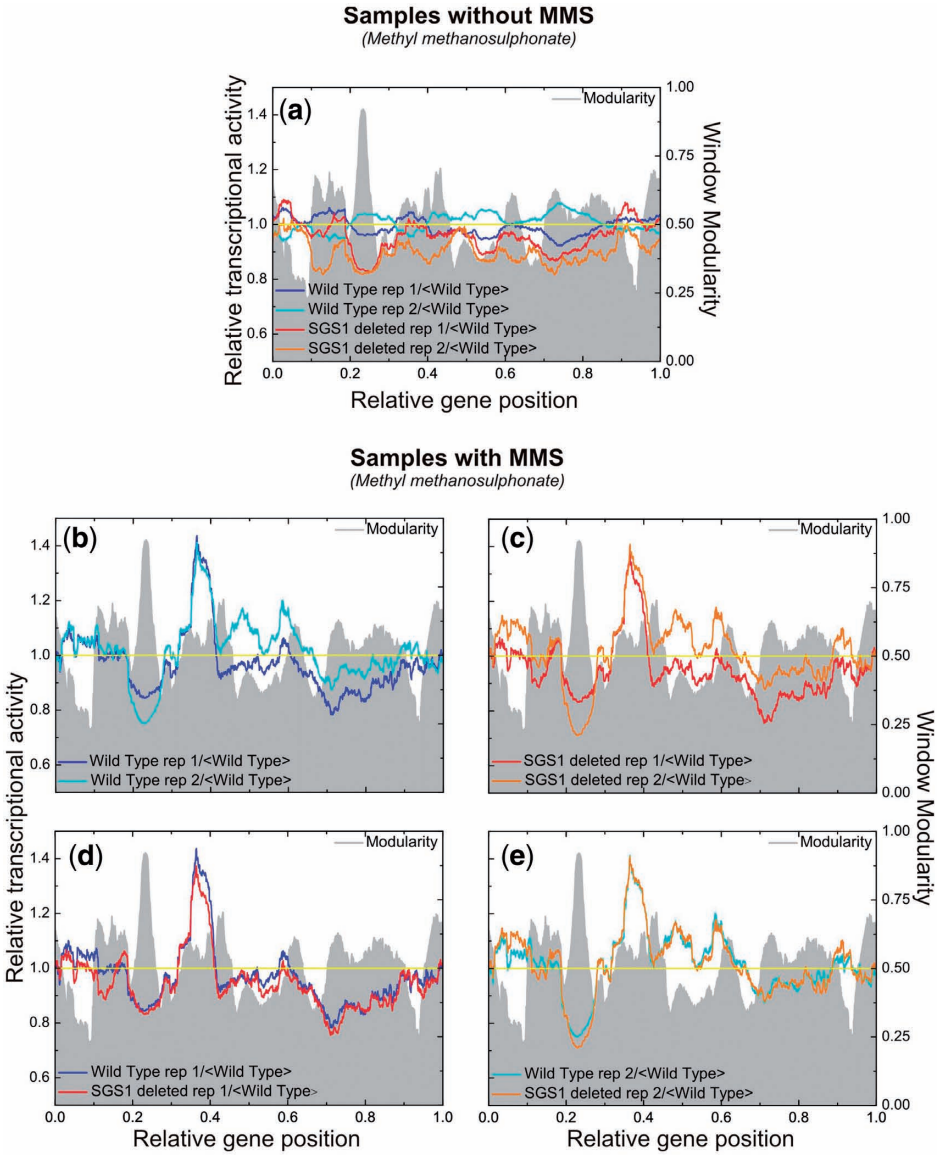


Figure 7. *Saccharomyces cerevisiae* transcriptograms: wild-type and sgs1 mutant. The axes relative to gene position have been divided by the total number of genes: 4655. (**a**) Transcriptograms for two replicates of wild-type and sgs1 mutant prior to addition of MMS. (**b**) Transcriptograms for two replicates of wild-type after MMS treatment. (**c**) Transcriptograms for two replicates of sgs1 mutants after MMS treatment. (**d**) Transcriptograms of one replicate of wild-type and one replicate of sgs1 mutant, to evince that both samples have been arrested at the same state of cell cycle. (**f**) Same as (**d**), but for the other pair of replicates, that have been arrested at a different state of cell cycle. All transcriptograms are taken relative to the average values of the replicates of wild-type, prior to the addition of MMS. Averages of windows of 251 have been taken.

presents the transcriptograms for their samples, always having the modularity as a background, to guide the eye. In those figures we have considered the expression levels relative to the average of each transcript of the wild-type under normal conditions.

Fugure 7a presents the transcriptogram for the normal, control condition for both wild-type and sgs1 mutants, in two replicates each, as obtained from Gene Expression Omnibus, GSE423, related to the experiment by Fry *et al*. We first observe that although there is not a very large peak, the transcriptogram of sgs1 mutants present an overall depression as compared to the wild-type replicate: sgs1 mutant relative transcription levels are consistently below the wild-type ones, possibly indicating a generalized reduction of cellular metabolism due to the knockout of sgs1 gene. Figure 7b and c present the transcriptograms for, respectively, wild-type and sgs1 mutants in two replicates each, after the treatment with MMS. The transcription levels were again taken in relation to the wild-type levels under normal conditions. Observe that each one of the figures present two very different transcriptograms. These differences are noticeable due to peaks and depressions as compared to the wild-type. However, taking into account the transcriptograms for respiration cycle presented in Figures 5 and 6, we can assume that in each case the replicates were arrested in different stages of the respiration cycle. In fact, addition of MMS can cause cell arrest in different stages of cell cycle (17). To evince further, Figure 7d and e present the superposition of transcriptograms of one replicate of the wild-type and one replicate of the sgs1 Mutant: the plots are now almost identical. This corroborates Fry and collaborators conclusions that, under MMS stress, the sgs1 mutant performs as the wild-type. However, it also indicates that care should be taken in what regards the cell cycle stage, by either synchronizing cell cycle stages as done by Tu *et al*. or assessing in which stage the cells are at the moment of measuring the transcription levels.

DISCUSSION

In summary, we propose here the transcriptogram as a tool for assessing cell metabolism, which is capable of discriminating the stage the cell is going through at a given instant, as well as pointing metabolic changes in altered cellular states as compared to a control state. The transcriptogram is capable of evincing these features due to the gene ordering that correlates the distance between any two genes in the ordering with the probability that they interact. Since for the ordering obtained using the CFM method this interaction probability decays exponentially with the distance between the genes, the neighborhood on the ordering may be used to obtain averages that smooth out too wild fluctuations presented by gene expression data. This correlation also allows the identification of different regions of the ordering with well defined metabolic functions, endowing the dynamics of expression levels with biological meaning. Furthermore, transcriptograms allow whole genome assessment of expression data, dispensing the clustering genes by their (maximally) altered expression levels, which may vary from a cell metabolic state to another.

Dendogram-like methods are capable of ordering the genome and may also be used to produce transcriptograms. However, they are less efficient in ordering at long-range neighborhood, and hence compromise the quality of the information evinced by the averages over neighboring sites, besides rendering more difficult the biological interpretation of the expression levels variations.

Further improvements on the algorithm should specifically consider window size, which ultimately reflects the functional correlation between genes. In fact, the transcriptogram opens the possibility of a tool that works as a telescope, where the focus is tuneable and may be adjusted to the desired level of detail: when passing from a wide genome overview to smaller functional modules analysis, the observation window may be narrowed down, discriminating more functional modules at greater detail. In this case, projecting smaller sets of functionally related genes as some KEGG pathways (18) may bring further information. On the other hand, the method is readily applicable to any species, including *Homo sapiens*, which will be presented elsewhere.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge fruitful discussions with Prof. Diego Bonatto, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul.

FUNDING

Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, partially); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, partially); Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS, partially). Funding for open access charge: Brazilian agencies CNPq.

Conflict of interest statement. None declared.

REFERENCES

1. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*., **30**, 1575–1584.
2. Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.
3. Blatt,M., Wiseman,S. and Domany,E. (1996) Superparamagnetic Clustering of Data. *Phy. Rev. Letts*., **76**, 3251.
4. Bader,G. and Hogue,C. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
5. Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barabasi,A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

6. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–115.
7. Metropolis,N. and Ulam,S. (1949) The Monte Carlo Method. *J. Amer. Stat. Assoc.*, **44**, 335–341.
8. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
9. The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–9.
10. Huang,d.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
11. Tu,B.P., Kudlicki,A., Rowicka,M. and McKnight,S.L. (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, **310**, 1152–1158.
12. Fry,R., Sambandan,T. and Rha,C. (2003) DNA damage and stress transcripts in Saccharomyces cerevisiae Mutant sgs1. *Mech. aging Dev.*, **124**, 839–846.
13. Kirkpatrick,S., Gelatt,C.D. and Vecchi,M.P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.
14. Vinogradov,A.E. (2008) Modularity of cellular networks shows general center-periphery polarization. *Bioinformatics*, **24**, 2814–2817.
15. Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.
16. Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.
17. Jelinsky,S.A., Estep,P., Church,G.M. and Samson,L.D. (2000) Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol. Cell. Biol.*, **20**, 8157–8167.
18. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,M., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Apêndice B

Artigo ViaComplex: software for landscape analysis of gene expression networks in genomic context

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 11 2009, pages 1468–1469
doi:10.1093/bioinformatics/btp246

Systems biology

ViaComplex: software for landscape analysis of gene expression networks in genomic context

Mauro A. A. Castro[1,*], José L. Rybarczyk Filho[2], Rodrigo J. S. Dalmolin[1], Marialva Sinigaglia[3], José C. F. Moreira[1], José C. M. Mombach[4] and Rita M. C. de Almeida[2]

[1]Unidade de Bioinformática, Departamento de Bioquímica, [2]Instituto de Física, [3]Núcleo de Bioinformática, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600-anexo, Porto Alegre 90035-003 and [4]Departamento de Física, Universidade Federal de Santa Maria, Santa Maria 97105-900, Brazil

Received on January 13, 2009; revised on March 22, 2009; accepted on April 6, 2009

Advance Access publication April 15, 2009

Associate editor: Thomas Lengauer

ABSTRACT

ViaComplex is an open-source application that builds landscape maps of gene expression networks. The motivation for this software comes from two previous publications (*Nucleic Acids Res*., **35**, 1859–1867, 2007; *Nucleic Acids Res.*, **36**, 6269–6283, 2008). The first article presents a network-based model of genome stability pathways where we defined a set of genes that characterizes each genetic system. In the second article we analyzed this model by projecting functional information from several experiments onto the gene network topology. In order to systematize the methods developed in these articles, ViaComplex provides tools that may help potential users to assess different high-throughput experiments in the context of six core genome maintenance mechanisms. This model illustrates how different gene networks can be analyzed by the same algorithm.

Availability: http://lief.if.ufrgs.br/pub/biosoftwares/viacomplex

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: mauro@ufrgs.br or rita@if.ufrgs.br

1 INTRODUCTION

Genome maintenance mechanisms (GMM) are critical for cell homeostasis. Evolution has shaped sophisticated repair systems that cover most of the insults that can cause genome damages. Defects in any of these systems can predispose to cancer (Castro *et al.*, 2008). At least four DNA damage repair pathways operate in mammals that, together with apoptosis and chromosome stability pathways, comprise the basis of GMM (Hoeijmakers, 2001; Zhivotovsky and Kroemer, 2004).

We have previously constructed a network-based model of human GMM in which different gene activity data were projected onto the interaction map (Castro *et al.*, 2007, 2008). Here, we extend these previous studies and develop a new software which could serve as a generalized tool to evaluate gene expression networks. With a graphical user interface, ViaComplex can either compute gene activity data for the internal model or import customized models of gene/protein interaction networks. In this case, the GMM network model illustrates the type of problem that can be dealt with the software for different gene networks. It can be used to produce publication quality images where data are visualized as functional landscapes projected onto gene network maps. ViaComplex also provides a statistical module based on the concept of information theory where multiple hypotheses are controlled by the false discovery rate (FDR) approach.

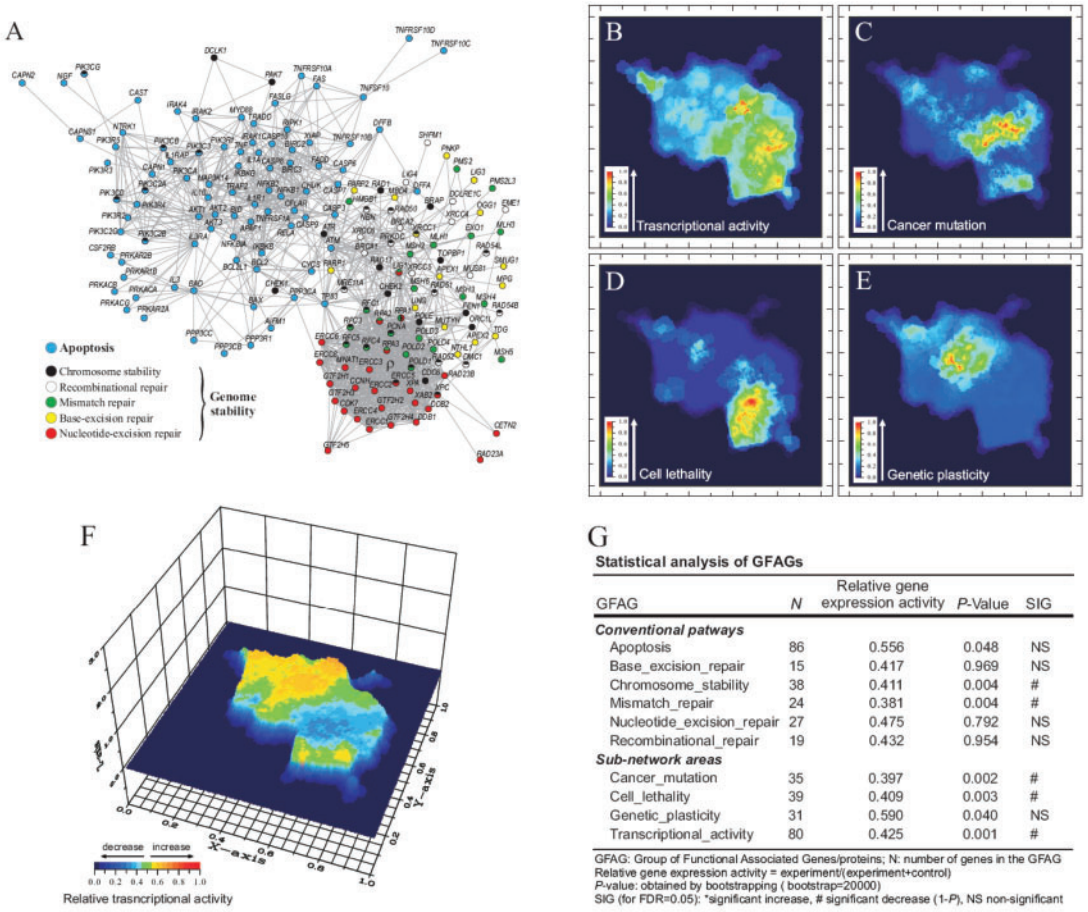
2 IMPLEMENTATION

ViaComplex program code is written for Linux and Windows Intel FORTRAN compilers (version 10.1.025) and is linked with Dislin 9.4, a scientific plotting library (http://www.dislin.de/). The main advantage of this program is that it is able to distribute a given quantity (quantitative or qualitative data) onto gene/protein interaction networks. To do this, ViaComplex overlaps functional information with interaction information (e.g. the network-based model of GMM).

The GMM network model comprises 180 genes that participate in human apoptosis and genome-stability functions as previously described (Castro *et al.*, 2007) and is depicted in Figure 1A. As an example of ViaComplex capabilities, in Figure 1B we show a microarray data analysis processed by the landscape module where gene expression activity is plotted over the network topology. In this figure the software distributed the microarray signal according to the coordinates of the network objects (i.e. nodes and links). By default, ViaComplex will distribute the signal on both nodes and links, but the user can change this option together with other ones available in the console. Also, the same algorithm can map qualitative data, as exemplified with cancer mutations (Fig. 1C), cell lethality (Fig. 1D) and genetic plasticity (Fig. 1E). Alternatively, user can compare two different functional states of the same gene network topology (Fig. 1F, methylated versus nonmethylated states).

The install package includes a comprehensive help file that provides the user all necessary details to prepare the data input and to execute the data analysis. ViaComplex can read the common

*To whom correspondence should be addressed.



GFAG	N	Relative gene expression activity	P-Value	SIG
Conventional patways				
Apoptosis	86	0.556	0.048	NS
Base_excision_repair	15	0.417	0.969	NS
Chromosome_stability	38	0.411	0.004	#
Mismatch_repair	24	0.381	0.004	#
Nucleotide_excision_repair	27	0.475	0.792	NS
Recombinational_repair	19	0.432	0.954	NS
Sub-network areas				
Cancer_mutation	35	0.397	0.002	#
Cell_lethality	39	0.409	0.003	#
Genetic_plasticity	31	0.590	0.040	NS
Transcriptional_activity	80	0.425	0.001	#

GFAG: Group of Functional Associated Genes/proteins; N: number of genes in the GFAG
Relative gene expression activity = experiment/(experiment+control)
P-value: obtained by bootstrapping (bootstrap=20000)
SIG (for FDR=0.05): *significant increase, # significant decrease (1-*P*), NS non-significant

Fig. 1. Landscape analysis of GMM. (**A**) Graph of interactions among genes involved GMM, as previously described in Castro *et al.* (2007). (**B**) Example of gene expression data analysis using breast MCF7 cell line (GEO accession no. GSM155194). Color gradient represents the transcriptional activity mapped onto graph. The same algorithm can be used to map different data, e.g. (**C**) genes causally implicated in human cancer, (**D**) yeast lethality data and (**E**) genetic plasticity, as defined in Castro *et al.* (2008). (**F**) Two-state landscape analysis. It compares the transcription profile of MCF7 cells in hypomethylated state (e.g. *state a*) versus hipermethylated state (e.g. *state b*) (GEO accessions GSE5816 versus GSM155194), where the colour gradient $Z = a/(a+b)$. A summary of the statistical analysis of this data is presented in (**G**). Figure 1A reprinted/adapted with permission from *NAR* (Oxford University Press).

gene/protein identifiers (e.g. EMBL, ENTREZ, UniProt, HGNC, RefSeq and UniGene) and the resulting graphs can be previewed as XWIN format or saved as PDF, EPS or PostScript files.

The install package also comprises an extensive library of published studies that exemplifies all procedures by a simple mouse click. In this sense, the GMM network model can be used to observe the functionality of the algorithm, which can analyze different gene networks. Supplementary Figure 1 illustrates this possibility for a large network (with 1892 genes). Such option is available at the 'custom model' module. It is semantically focused in genes, matching gene IDs and node IDs. If there is more than one microarray probe interrogating the expression of a given gene, then the software takes the average of the expression values, which allows the use of different microarray platforms (i.e. it does not involve comparisons between network nodes and probe tag IDs). Other numerical samples of different sizes are available at ViaComplex homepage.

Additional features of the software include a statistical module where two microarray datasets can be compared following the method described in Castro *et al.* (2007), as exemplified in Figure 1G for the data used to build the Figure 1F. We anticipate that ViaComplex will be useful to mine graph patterns from high-throughput experimental data.

Funding: Brazilian Agencies FAPERGS, CAPES and CNPq.

Conflict of Interest: none declared.

REFERENCES

Castro,M.A.A. *et al.* (2007) Impaired expression of NER gene network in sporadic solid tumors. *Nucleic Acids Res.*, **35**, 1859–1867.

Castro,M.A.A. *et al.* (2008) Evolutionary origins of human apoptosis and genome-stability gene networks. *Nucleic Acids Res.*, **36**, 6269–6283.

Hoeijmakers,J.H.J. (2001) Genome maintenance mechanisms for preventing cancer. *Nature*, **411**, 366–374.

Zhivotovsky,B. and Kroemer,G. (2004) Apoptosis and genomic instability. *Nat. Rev. Mol. Cell Biol.*, **5**, 752–762.

Referências Bibliográficas

[AND 08] ANDRADE, R. F. et al. Measuring distances between complex networks. **Physics Letters A**, New York, USA, v.372, n.32, p.5265 – 5269, 2008.

[ASH 94] ASHBURNER, M.; DRYSDALE, R. Flybase–the drosophila genetic database. **Development**, Hampshire, USA, v.120, n.7, p.2077–2079, Jul, 1994.

[ASH 00] 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.

[BAD 03] BADER, G.; HOGUE, C. An automated method for finding molecular complexes in large protein interaction networks. **BMC Bioinformatics**, London, UK, v.4, n.1, p.2+, January, 2003.

[BAR 01] 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.

[BAR 03] BARABáSI, A.-L.; BONABEAU, E. Scale-free networks. **Sci Am**, London, UK, v.288, n.5, p.60–69, May, 2003.

[BAR 07] 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.

[CAM 65] 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.

[CAS 09] CASTRO, M. A. A. et al. Viacomplex: software for landscape analysis of gene expression networks in genomic context. **Bioinformatics**, Oxford, UK, v.25, n.11, p.1468–1469, Jun, 2009.

[CHE 98] CHERRY, J. M. et al. Sgd: Saccharomyces genome database. **Nucleic Acids Res**, Oxford, UK, v.26, n.1, p.73–79, Jan, 1998.

[CRA 89] CRAWFORD, M. H. Hugo: Genome data open to scientists. **Science**, Washington, USA, v.246, n.4937, p.1565, Dec, 1989.

[DEE 06] DEEDS, E. J.; ASHENBERG, O.; SHAKHNOVICH, E. I. A simple physical model for scaling in protein-protein interaction networks. **Proc Natl Acad Sci U S A**, Washington, USA, v.103, n.2, p.311–316, Jan, 2006.

[dN 05] DE NOOY, W.; MRVAR, A.; BATAGELJ, V. **Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences)**. Cambridge, UK: Cambridge University Press, January, 2005.

[EDG 02] 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.

[ENR 02] ENRIGHT, A. J.; DONGEN, S. V.; OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. **Nucleic Acids Res**, Oxford, UK, v.30, n.7, p.1575–1584, Apr, 2002.

[ERD 60] 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.

[FEL 97] FELSENSTEIN, J. An alternating least squares approach to inferring phylogenies from pairwise distances. **Syst Biol**, London, UK, v.46, n.1, p.101–111, Mar, 1997.

[FRA 05] FRASER, H. B. Modularity and evolutionary constraint on proteins. **Nat Genet**, London, UK, v.37, n.4, p.351–352, Apr, 2005.

[FRU 91] FRUCHTERMAN, T. M. J.; REINGOLD, E. M. Graph drawing by force-directed placement. **Softw. Pract. Exper.**, New York, NY, USA, v.21, n.11, p.1129–1164, 1991.

[FRY 03] FRY, R. C.; SAMBANDAN, T. G.; RHA, C. Dna damage and stress transcripts in saccharomyces cerevisiae mutant sgs1. **Mech Ageing Dev**, Hampshire, USA, v.124, n.7, p.839–846, Jul, 2003.

[GER 95] GERSTING, J. L. **Fundamentos matemáticos para ciência da computação.** 3º. ed. São Paulo, Brasil: LTC, 1995.

[GOH 07] 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.

[GU 06] GU, Q.; SIVANANDAM, T. M.; KIM, C. A. Signal stability of cy3 and cy5 on antibody microarrays. **Proteome Sci**, London, UK, v.4, p.21, 2006.

[HAK 07] HAKES, L. et al. Specificity in protein interactions and its relationship with sequence diversity and coevolution. **Proc Natl Acad Sci U S A**, Washington, USA, v.104, n.19, p.7999–8004, May, 2007.

[HID 09] 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.

[HOO 05] HOOPER, S. D.; BORK, P. Medusa: a simple tool for interaction graph analysis. **Bioinformatics**, Oxford, UK, v.21, n.24, p.4432–4433, 2005.

[HUA 07] HUANG, D. W. et al. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. **Genome Biol**, London, UK, v.8, n.9, p.R183, 2007.

[HUA 09] HUANG, D. W.; SHERMAN, B. T.; LEMPICKI, R. A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. **Nat Protoc**, London, UK, v.4, n.1, p.44–57, 2009.

[JEL 00] JELINSKY, S. A. et al. Regulatory networks revealed by transcriptional profiling of damaged saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes. **Mol Cell Biol**, Washington, USA, v.20, n.21, p.8157–8167, Nov, 2000.

[JEN 09] 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.

[KAN 02] KANEHISA, M. The kegg database. **Novartis Found Symp**, New York, USA, v.247, p.91–101; discussion 101–3, 119–28, 244–52, 2002.

[LOS 07] LOSCALZO, J.; KOHANE, I.; BARABASI, A.-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. **Mol Syst Biol**, London, UK, v.3, p.124, 2007.

[MCL 08] MCLACHLAN, G. J.; BEAN, R. W.; NG, S.-K. Clustering. **Methods Mol Biol**, New York, USA, v.453, p.423–439, 2008.

[NEL 04] NELSON, D. L.; COX, M. M. **Lehninger Principles of Biochemistry, Fourth Edition**. Fourth Edition. ed. New York, USA, 2004.

[NEW 06] 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.

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

[OKU 08] OKUDA, S. et al. Kegg atlas mapping for global analysis of metabolic pathways. **Nucleic Acids Res**, Oxford, UK, v.36, n.Web Server issue, p.W423–W426, Jul, 2008.

[OLT 02] OLTVAI, Z. N.; BARABáSI, A.-L. Systems biology. life's complexity pyramid. **Science**, Washington, USA, v.298, n.5594, p.763–764, Oct, 2002.

[OLT 04] OLTVAI, A.-L. B. . Z. N. Network biology: understanding the cell's functional organization. **Nature Reviews Genetics**, London, UK, v.5, p.101–113, 2004.

[PO 01] PAULO OSWALDO, B. N. **Grafos: teorias,modelos e algoritmos.** 4ª. ed. São Paulo, Brasil: Edgar Blücher, 2001. 328 p.

[RA 08] R.F.S. ANDRADE, J.G.V. MIRANDA, S. P.; LOBãO, T. Characterization of complex networks by higher order neighborhood properties. **The European Physical Journal B**, Les Ulis, France, v.61, p.247–256, 2008.

[RAV 02] RAVASZ, E. et al. Hierarchical organization of modularity in metabolic networks. **Science**, Washington, USA, v.297, n.5586, p.1551–1555, Aug, 2002.

[RAV 03] 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.

[RF 10] RYBARCZYK-FILHO, J. L. et al. Towards a genome-wide transcriptogram: the saccharomyces cerevisiae case. **Nucleic Acids Res**, Oxford, UK, Dec, 2010.

[SAI 04] SAID, M. R. et al. Global network analysis of phenotypic effects: protein networks and toxicity modulation in saccharomyces cerevisiae. **Proc Natl Acad Sci U S A**, Washington, USA, v.101, n.52, p.18006–18011, Dec, 2004.

[SPI 03] SPIRIN, V.; MIRNY, L. A. Protein complexes and functional modules in molecular networks. **Proc Natl Acad Sci U S A**, Washington, USA, v.100, n.21, p.12123–12128, Oct, 2003.

[SPI 06] SPIRIN, V. et al. A metabolic network in the evolutionary context: multiscale structure and modularity. **Proc Natl Acad Sci U S A**, Washington, USA, v.103, n.23, p.8774–8779, Jun, 2006.

[STR 03] STRONG, M. et al. Visualization and interpretation of protein networks in mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps. **Nucleic Acids Res**, Oxford, UK, v.31, n.24, p.7099–7109, Dec, 2003.

[SWA 08] SWARBRECK, D. et al. The arabidopsis information resource (tair): gene structure and function annotation. **Nucleic Acids Res**, Oxford, UK, v.36, n.Database issue, p.D1009–D1014, Jan, 2008.

[TOY 07] TOYODA, T. et al. Omicbrowse: a browser of multidimensional omics annotations. **Bioinformatics**, Oxford, UK, v.23, n.4, p.524–526, 2007.

[TU 05] TU, B. P. et al. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. **Science**, Washington, USA, v.310, n.5751, p.1152–1158, Nov, 2005.

[VIN 08] VINOGRADOV, A. E. Modularity of cellular networks shows general center-periphery polarization. **Bioinformatics**, Oxford, UK, v.24, n.24, p.2814–2817, Dec, 2008.

[vM 05] 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.

[vM 07] 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.

[WAT 53a] 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.

[WAT 53b] 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.

[YOO 04] YOOK, S.-H.; OLTVAI, Z. N.; BARABáSI, A.-L. Functional and topological characterization of protein interaction networks. **Proteomics**, Malden, USA, v.4, n.4, p.928–942, Apr, 2004.