UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE FÍSICA
Trabalho de Conclusão de Curso

Homo sapiens: análise de expressão gênica por transcriptograma

Fernanda Pereira da Cruz Benetti

Trabalho de conclusão de curso, realizado sob orientação da professora Dra. Rita Maria Cunha de Almeida, apresentado ao Instituto de Física da UFRGS em preenchimento parcial dos requisitos para a obtenção do título Bacharel em Física.

Porto Alegre, novembro de 2010.

Resumo

O proteoma – o grupo de proteínas produzidas na célula – e suas interações formam uma rede complexa que determina o funcionamento e desenvolvimento celular. Devido à enorme quantidade de dados disponíveis, novas ferramentas são necessárias para organizar e identificar processos relevantes dentro do proteoma, permitindo uma análise de larga-escala (i.e. proteômica) em vez de pequena-escala (i.e. protéica). Neste trabalho, usamos um algoritmo de minimização de custo para organizar o proteoma humano e identificar módulos interativos, e mostramos que esses módulos podem ser compostos por proteínas de funções biológicas semelhantes. Finalmente, apresentamos o transcriptograma de células de tecido pulmonar humano normal e canceroso (adenocarcinoma). O transcriptograma mostra níveis de expressão gênica ao longo da rede ordenada, possibilitando a análise de larga-escala da dinâmica celular e tornando-o uma ferramenta poderosa de diagnóstico. Neste estudo, mostramos que ele pode ser usado para identificar módulos de proteínas que são expressas em níveis alterados em tecido normal e canceroso de fumantes e ex-fumantes, em comparação com tecido normal de não-fumantes.

Abstract

The proteome – the group of proteins produced in the cell – and their interactions form a complex network that determines the cell's functioning and development. Due to the enormous amount of data available, new methods and tools are needed to sort through and organize relevant processes within the proteome so as to allow a large-scale (i.e. proteomic) rather than small-scale (i.e. proteic) analysis. Here we use a cost-minimization algorithm to organize the human proteome and identify interactive modules, and show that these modules can be composed of proteins of similar biological functions. Finally, we present the transcriptogram of normal and cancerous (adenocarcinoma) cells of human lung tissue. The transcriptogram shows the gene expression levels along the ordered network, thereby enabling the large-scale analysis of the cell's dynamics and making it a powerful tool for diagnosis. In this study, we show that it can be used to identify modules of proteins that are expressed at altered levels in both normal and cancerous tissue of smokers and ex-smokers, in comparison to normal tissue of non-smokers.

Sumário

1 Introdução **2**
1.1 Objetivos . . . 3
1.2 Resumo sobre redes . . . 4
1.3 Expressão gênica e proteínas . . . 5

2 Método **6**
2.1 A rede protéica e sua modelagem . . . 6
2.2 Ordenamento . . . 7
2.3 Modularidade por janela . . . 11
2.4 Atribuição de funções celulares . . . 11
2.5 Análise de expressão gênica . . . 12
2.6 Diferenças entre nomenclaturas . . . 13

3 Resultados **14**
3.1 Ordenamento final . . . 14
3.2 Módulos . . . 14
3.3 Associação de processos biológicos . . . 16
3.4 Transcriptograma . . . 17

4 Conclusões **29**

Referências Bibliográficas **32**

Capítulo 1

Introdução

Os avanços do último século na compreensão da origem e do funcionamento básico dos organismos foram possíveis graças à descoberta dos genes e do DNA. Todo organismo, seja ele um microorganismo unicelular como uma bactéria ou um ser composto por trilhões de células como o ser humano, tem seu desenvolvimento e funcionamento regido pela informação contida no DNA. Essa macromolécula contém as instruções para a célula: genes, que são segmentos de DNA, são lidos e transcritos por moléculas de RNA que, por sua vez, são traduzidas por ribossomos resultando na produção de proteínas. Essa área de pesquisa apresentou grandes desafios como o deciframento do código genético e o sequênciamento do genoma (o conjunto de genes de um organismo). No entanto, revelou-se insuficiente conhecer o código e catalogar os genes e as proteínas codificadas por eles, devido à complexidade do sistema que eles formam.

Tanto os genes quanto as proteínas interagem entre si. A influência de uma proteína sobre um organismo só é adequadamente compreendida ao examinar sua relação com outras proteínas. O conjunto total de proteínas codificadas pelos genes de um organismo é conhecido como *proteoma*, e as interações proteína-proteína e proteína-gene formam o *interatoma*. A *rede protéica* é o conjunto de proteínas e suas interações. É um sistema complexo pois envolve muitos componentes que interagem entre si. Uma mudança num componente da rede pode alterar drasticamente seu comportamento ou pode não ter influência alguma, dependendo do tipo e quantidade de interações em que ele participa.

Uma questão importante no estudo de redes protéicas é se existe ou não um padrão topológico entre as redes de diferentes organismos. Semelhanças entre redes protéicas podem denotar proximidade na árvore evolutiva, já que o proteoma é consequência direta do genoma, que por sua vez é o sistema fundamentalmente afetado por pressões evolutivas. Outros trabalhos já demonstraram que há características compartilhadas por proteomas de diferentes organismos, como a modularidade [1, 2]. Uma rede modular é aquela que pode ser decomposta em subredes, ou módulos, de elementos que interagem muito mais entre si do que com aqueles externos ao grupo.

A enorme quantidade de informação sobre genes, proteínas e suas interações torna a

classificação de conjuntos de proteínas como módulos uma tarefa não-trivial. O algoritmo de minimização de custo [3] organiza a rede protéica, aglomerando as proteínas que mais interagem entre si. Dessa forma torna-se possível identificar módulos.

A identificação por si é interessante pois torna evidente a estrutura modular intrínseca das redes protéicas, mostrando claramente que o proteoma não corresponde a uma rede de construção aleatória (ver seção 1.2). Por outro lado, o aspecto mais vantajoso da identificação é a possibilidade de associar aos módulos funções celulares distintas, demonstrando que a estrutura modular do proteoma corresponde a um funcionamento modular da célula.

Outro aspecto do sistema celular a ser analisado é a sua dinâmica. Enquanto a classificação funcional atribui uma característica à proteína que é invariante no tempo, o nível de expressão gênica contém informação sobre o estado e as condições em que a célula se encontra naquele momento. Uma análise global da transcrição (uma etapa da expressão) mostra quais genes estão ativos, ou que proteínas estão sendo produzidas, no instante em que a medida de expressão foi realizada. O *transcriptograma* [3] é a análise de transcrição sobre a rede protéica ordenada. Dessa maneira, pode-se estudar o efeito de diferentes condições (internas ou ambientais) sobre a atividade transcricional dos módulos que, por sua vez, representam funções celulares. O transcriptograma é uma nova ferramenta de diagnóstico do funcionamento celular.

Em seu artigo, Rybarczyk Filho et al usaram o algoritmo para ordenar e fazer transcriptogramas do genoma de *Saccharomyces cerevisiae* [3]. Nesse trabalho, aplicamos o mesmo método para o genoma de *Homo sapiens*. Como uma primeira aplicação do transcriptograma para fins diagnósticos em humanos, analisamos células de tecido pulmonar canceroso.

As próximas seções desse capítulo expõem os objetivos do trabalho e explicam resumidamente alguns conceitos de redes, grafos e expressão gênica. Nos capítulos seguintes, primeiramente abordamos o método de ordenamento, as medidas usadas para estudar o proteoma e a realização do transcriptograma. Então apresentamos os resultados do ordenamento, os módulos identificados e os transcriptogramas de tecido humano normal (não-tumoral) e canceroso. O último capítulo mostra as interpretações dos resultados, as conclusões e as perspectivas para continuar e aprofundar essa pesquisa.

1.1 Objetivos

Esse trabalho tem três objetivos principais, baseados nas motivações já expostas:

1. Organizar a rede protéica do *Homo sapiens* de forma a evidenciar módulos distintos.

2. Identificar processos biológicos que sejam representativos de cada módulo.

3. Usar o transcriptograma para analisar e comparar a atividade transcricional de células cancerosas e células saudáveis de fumantes, ex-fumantes e não-fumantes diagnosticados com câncer pulmonar.

Visando o melhor entendimento tanto do algoritmo quanto da análise de transcrição, alguns conceitos básicos de redes e de expressão gênica devem ser explicados.

1.2 Resumo sobre redes

Redes são sistemas formados por *nós* e *ligações* (*vértices* e *arestas*, na terminologia da teoria de grafos). Os nós podem representar quaisquer tipos de elementos que interagem entre si. Uma interação é representada pela ligação que conecta um nó a outro. Redes podem ser direcionadas ou não-direcionadas: se a ligação de um nó A a outro nó B não equivale a uma ligação de B a A, a rede é direcionada. Se as ligações forem equivalentes, é não-direcionada [4].

Diferentes tipos de rede podem ser classificadas de acordo com alguns parâmetros como seu grau de conectividade e clusterização, entre outros. O grau ou conectividade k_i de um nó i é o número de nós aos quais ele está ligado (seus vizinhos na rede). A conectividade média $< k >$ da rede é a média aritmética da conectividade dos nós. Numa rede não-direcionada de N nós e L ligações, $< k >= \frac{2L}{N}$ [4, 2].

O coeficiente de clusterização C_i quantifica a interatividade entre os vizinhos do nó i: é a razão entre o número de ligações entre os vizinhos de i e o número de possíveis ligações entre eles (se todos estivessem conectados), e pode ser expresso como [2]

$$C_i = \frac{2l_i}{k_i(k_i - 1)}, \tag{1.1}$$

onde l_i é o número de ligações entre os k_i vizinhos do nó i.

Outras medidas importantes da rede são a distribuição de conectividade $P(k)$, que é a probabilidade de dado nó ter conectividade k, e a distribuição de clusterização $C(k)$, que é a clusterização média dos nós de conectividade k. Uma rede construída de forma aleatória (a probabilidade de que um nó inserido na rede se conecte ao nó i é igual para todo i) terá como $P(k)$ uma distribuição normal. A rede aleatória não é modular, pois a maioria dos nós terá aproximadamente o mesmo número de ligações e não há grupos que estejam mais ligados entre si. Existem outros modelos de rede que também não são modulares, como a *scale-free* [5]. Não está no escopo desse trabalho aprofundar esses modelos, mas apenas ressaltar que existem diferentes modelos topológicos de redes, que podem ser caracterizadas e diferenciadas por parâmetros independentes do tamanho da rede, como $P(k)$ e $C(k)$ [2].

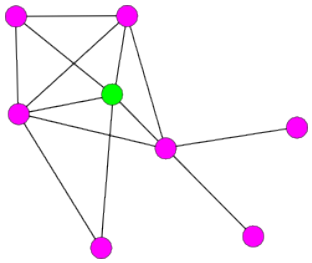

Figura 1.1: Exemplo de rede. Em relação ao nó verde: $k = 5$, $l = 6$ e $C = \frac{2l}{k(k-1)} = \frac{3}{5}$.

1.3 Expressão gênica e proteínas

Cada gene contém instruções para a produção de alguma molécula dentro da célula. Para que isso ocorra, a informação contida nele deve ser lida, traduzida e utilizada para sintetizar seu produto final. Esse processo se denomina expressão gênica: é a expressão da informação codificada no gene. Um gene é um segmento de DNA que codifica uma cadeia polipeptídica ou de RNA. O tipo de expressão gênica tratada nessa seção é aquela que resulta na síntese de proteínas (uma ou mais cadeias polipeptídicas), que é a mais comum [6].

Expressão gênica ocorre em várias etapas, das quais as duas principais são a transcrição e a tradução. Transcrição consiste em replicar a informação contida no gene na forma de uma molécula de RNA. A enzima RNA-polimerase separa as duas cadeias que formam a molécula de DNA e combina bases nitrogenadas de RNA às bases nitrogenadas complementares do DNA. Por exemplo, uma sequência ATCG no DNA resulta numa sequência UAGC no RNA (a timina [T] é substituída pela uracila [U] no RNA). Após a transcrição, as ligações entre o RNA e o DNA se rompem e, no caso da síntese protéica, a molécula de RNA é transportada para o citoplasma (no caso de eucariotos) onde ela será traduzida. Nesse caso, ela é denominada mRNA, ou RNA mensageiro. A tradução do mRNA é conduzida por ribossomos. O ribossomo liga bases de mRNA a bases de tRNA (RNA de transporte) que por sua vez estão ligadas a aminoácidos. A cadeia formada pelos aminoácidos é a proteína, o resultado final da expressão. Em alguns casos, várias cadeias de aminoácidos (cada uma codificada por um gene diferente) formam uma única proteína. Em função disso, uma proteína pode ser codificada por mais de um gene, e um gene pode ser responsável pela formação de mais de um tipo de proteína [6].

A síntese protéica é fundamental para o organismo, pois quase todo processo biológico é realizado por proteínas. Existem muitos tipos diferentes de proteínas dentro de uma única célula, cada uma exercendo uma função específica. Por exemplo: reações químicas entre biomoléculas são catalisadas por enzimas; moléculas e íons são carregadas entre membranas, órgãos e tecidos por proteínas de transporte; tecidos e fibras são compostos por proteínas estruturais; proteínas de defesa protegem contra invasores e lesões; etc [6]. Um dos objetivos do ordenamento obtido pela minimização de custo é separar grupos bem-definidos de proteínas interativas que sejam responsáveis pelos mesmos processos biológicos, que não é trivial.

Capítulo 2

Método

O algoritmo consiste em usar a técnica de Monte Carlo para minimizar a função custo de uma matriz de interação da rede protéica. As informações sobre a rede (lista das proteínas e suas interações) são obtidas do STRING [7], uma base de dados disponível online. Nesse capítulo, primeiramente explica-se a natureza do tipo de interações consideradas para formar a rede. Em seguida, expomos o método de ordenamento, as medidas usadas para caracterizar a rede e, finalmente, o transcriptograma.

2.1 A rede protéica e sua modelagem

A rede protéica é o conjunto das proteínas codificadas pelos genes e das interações entre elas. É importante ressaltar que o termo *interação* utilizado nesse trabalho não se refere exclusivamente a interações proteína-proteína, que são interações físicas diretas, mas também a associações funcionais como a participação do mesmo processo celular ou da mesma rota metabólica. As informações sobre o interatoma humano foram obtidas através da base de dados STRING, versão 8.2. O STRING reúne e disponibiliza informações sobre interações físicas diretas (proteína-proteína) e indiretas (rotas metabólicas), além de interações que são previstas por diferentes métodos como:

Associação por vizinhança

A proximidade de genes no genoma, quando frequente, indica que as proteínas codificadas por eles participam da mesma rota metabólica. Nesse caso, considera-se que as proteínas interagem entre si.

Associação por co-ocorrência

Outra maneira de inferir associações funcionais entre proteínas é pela história evolutiva de seus genes correspondentes. O perfil filogenético de um gene mostra seu surgimento e desaparecimento nos genomas de diferentes espécies. Dependendo do grau de semelhança entre esses perfis, as proteínas são funcionalmente associadas.

Associação por fusão

Dois genes podem, como resultado de pressão seletiva, se transformar num único gene. Esse processo se chama fusão de genes e já foi mostrado que há correlação entre tal evento e associação funcional de proteínas.

Associação por co-expressão

O nível de atividade transcricional de um gene depende das condições em que a célula se encontra. Há associação de proteínas quando seus genes codificadores têm perfil de transcrição muito parecido.

Transferência de associações

Existem métodos para inferir se, no caso de duas proteínas interagirem num organismo, elas também interagem noutro organismo. Algumas interações são obtidas dessa maneira.

Text mining

Se duas proteínas são frequentemente citadas no mesmo texto científico, o STRING considera que elas interagem, dado que a co-ocorrência de citações seja estatisticamente relevante. Esse tipo de previsão não foi considerado para formar a rede protéica usada nesse trabalho.

No STRING, é possível selecionar quais os tipos de associações funcionais que são desejadas para formar a rede, além de filtrá-las por seu grau de confiança, ou *score* [8]. A rede utilizada nesse trabalho considera as associações citadas acima (fora o text mining) e tem um score combinado de $0,8$. O score combinado é uma expressão dos scores individuais de cada ligação, considerando-as independentes, e é dado por

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

onde S_i é o score do método de predição i.

Esse score indica a probabilidade de que as interações previstas pelo STRING realmente existam, ou seja, de que a maioria das ligações previstas é resultado verdadeiro-positivo. Por exemplo, uma ligação que é prevista por todos os tipos de associação listados acima é mais confiável do que uma que é prevista por apenas um tipo. Um score total alto ($> 0,7$) diminui o número de falsos-positivos (ligações previstas, porém inexistentes) na rede, mas tem a desvantagem de eliminar muitos falsos-negativos (ligações imprevistas, porém existentes). O tamanho da rede protéica usada nesse trabalho (9019 nós e 111602 ligações) se deve ao alto score utilizado.

2.2 Ordenamento

Para ordenar a rede, o algoritmo primeiramente ordena N nós, representando as N proteínas do proteoma, aleatoriamente, dentro do intervalo [1,N]. Seguindo este ordenamento, constrói-

se uma matriz de interaçoes tal que

$$M_{i,j} = \begin{cases} 1 & \text{se } i \text{ e } j \text{ têm ligação} \\ 0 & \text{se } i \text{ e } j \text{ não têm ligação.} \end{cases} \tag{2.2}$$

A rede é não-direcionada, pois considerar que uma proteína A é funcionalmente associada a uma proteína B é o mesmo que considerar que a B está associada a A. Portanto, a matriz de interação é simétrica: $M_{i,j} = M_{j,i}$.

A figura 2.1 mostra a rede usada como exemplo na seção 1.2 e sua matriz de interação, após a enumeração aleatória dos nós.

Posição	1	2	3	4	5	6	7	8
Nó	H	G	D	C	B	A	F	E

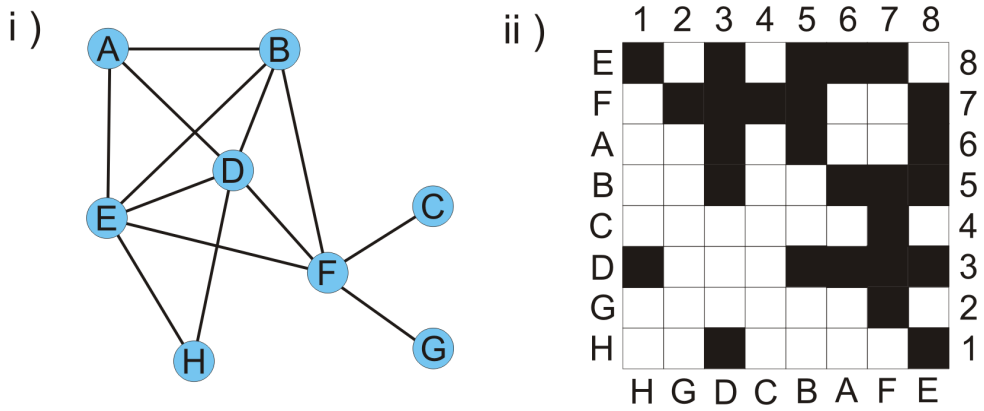

Figura 2.1: A rede exemplo e sua matriz de interação. Quadrados pretos indicam que há ligação ($M_{i,j} = 1$) e brancos indicam que não há ($M_{i,j} = 0$).

A distribuição inicial de valores 1 e 0 deve ser aproximadamente uniforme, pois os nós são aleatoriamente posicionados. O objetivo do ordenamento é usar técnicas de Monte Carlo para chegar a uma configuração na qual os nós que interagem muito entre si estejam próximos uns dos outros e longes dos demais. Isso corresponde a aproximar os sítios de valor 1 à diagonal da matriz. Como exemplo, tomemos dois nós A e B que têm ligação entre si. No ordenamento inicial, é provável que eles não sejam vizinhos, pois foram designadas posições aleatórias. Quanto maior a diferença entre suas posições, mais longe da diagonal ficará o sítio que representa sua ligação.

A l-ésima diagonal é o vetor $\mathbf{d}_l = \{M_{i,i+l}\}_{i=1}^{N-l}$. A probabilidade de haver interação entre duas proteínas dado que elas estão separadas por uma distância l no ordenamento é a densidade de pontos na l-ésima diagonal,

$$P\left(M_{i,j} = 1 || i - j| = l\right) = \frac{\sum_{i'=1}^{N-l} M_{i',i'+l}}{N - l}. \tag{2.3}$$

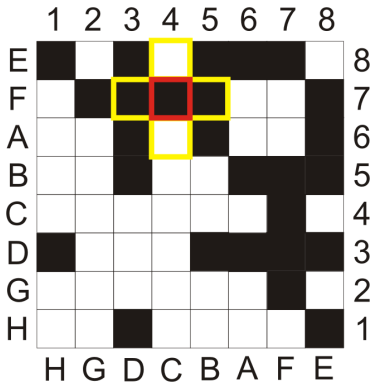

Figura 2.2: A contribuição de $M_{4,7}$ no ordenamento inicial da rede exemplo é $|4-7|(1+0+1+0)=6$. O sítio $(4,7)$ está destacado em vermelho e seus primeiros vizinhos em amarelo.

No ordenamento ideal, $P\left(M_{i,j}=1||i-j|=l\right)$ é máxima em $l=1$ e decai rapidamente à medida que l cresce.

Para atingir o objetivo de aproximar os nós mais interativos, foi determinada uma função de custo H para a matriz. Essa função penaliza *i*) a distância de sítios de valor 1 até a diagonal da matriz; e *ii*) o gradiente entre um sítio e seus primeiros vizinhos na matriz. Ela é definida como

$$H=\sum_{i=1}^{N}\sum_{j=1}^{N}|i-j|\left(|M_{i,j}-M_{i+1,j}|+|M_{i,j}-M_{i-1,j}|\right.$$
$$\left.+|M_{i,j}-M_{i,j+1}|+|M_{i,j}-M_{i,j-1}|\right), \tag{2.4}$$

onde observamos condições de contorno periódicas para as bordas da matriz. O valor absoluto da diferença entre i e j é proporcional à distância do sítio (i,j) até a diagonal e, nesse caso, as condições de contorno periódicas não se aplicam.

O primeiro passo do algoritmo é calcular a função de custo do estado inicial do sistema. Em seguida, dois números inteiros k_1 e k_2 são gerados, com probabilidade uniforme dentro do intervalo $[1,N]$. Os nós na posição k_1 e k_2 no ordenamento são trocados; consequentemente, a matriz de interação também é modificada — a coluna k_1 e a coluna k_2 são trocadas, assim como a linha k_1 e a linha k_2. Como a coluna/linha inteira muda de posição, nenhuma informação sobre as ligações dos nós k_1 e k_2 é perdida.

A função de custo da nova configuração, H', é calculada e comparada com o valor da configuração anterior, H. Se a diferença entre elas, $\Delta H=H'-H$, for negativa, a troca de posição dos nós é aceita. Caso contrário, ela é aceita com probabilidade $P=\exp\left(-\frac{\Delta H}{T}\right)$, na qual T, a temperatura, é um parâmetro usado para evitar estados metaestáveis. Após aceitar ou rejeitar a troca, o processo reinicia com o sorteio de dois novos números k_1 e k_2.

Em resumo, o algoritmo de minimização de custo é o seguinte:

1. Enumerar aleatoriamente os nós da rede para obter o ordenamento inicial e montar a matriz de interação M.

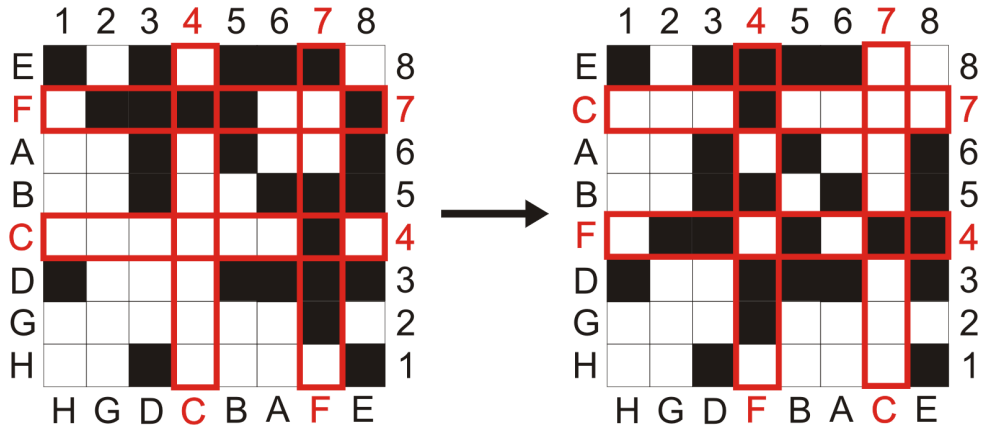

Figura 2.3: Troca de duas linhas e duas colunas da matriz de interação da rede exemplo ($k_1 = 4$, $k_2 = 7$). Essa troca é aceita com probabilidade 1 se $\Delta H < 0$ ou $\exp -\frac{\Delta H}{T}$ se $\Delta H > 0$.

2. Calcular a função custo inicial, H_0.

3. Sortear dois números, k_1 e k_2, trocar suas posições no ordenamento e montar a nova matriz de interação, M':

$$(1\,,2\,,\ldots\,,k_1\,,\ldots\,,k_2\,,\ldots\,,N) \to (1\,,2\,,\ldots\,,k_2\,,\ldots\,,k_1\,,\ldots\,,N)$$

$$M'_{k_{1,2}\,,j} = M_{k_{2,1}\,,j}, \quad j = 1,\ldots,N$$

$$M'_{i\,,k_{1,2}} = M_{i\,,k_{2,1}}, \quad i = 1,\ldots,N$$

4. Calcular $\Delta H = H' - H$, onde H' é o custo de M'.

5. Aceitar troca com probabilidade

$$P = \begin{cases} 1 & \text{se} \quad \Delta H \leq 0 \\ \exp\left(-\frac{\Delta H}{T}\right) & \text{se} \quad \Delta H > 0 \end{cases} \tag{2.5}$$

6. Voltar ao passo 3.

Esse processo é repetido até que sistema se estabilize numa configuração de mínimo custo. No entanto, pode haver muitas configurações de baixo custo e o sistema pode ficar preso num único estado. Nessa situação qualquer pequena alteração leva o sistema a uma configuração de maior custo e trocas são raramente aceitas. Isso é um problema se o estado em que ele se encontra for um mínimo local, o que é indesejado porque pode haver configurações de custo ainda menor que não são alcançadas. Para evitar que isso ocorra, a temperatura inicialmente é alta (da ordem de $H_0/100$) e constante durante um número fixo τ de passos de Monte Carlo; após esse tempo, ela é baixada por um fator γ e mantida fixa novamente durante o mesmo intervalo de passos. Isso é repetido até que ela seja quase nula no final. Com isso, o sistema inicialmente consegue explorar o espaço de possíveis configurações, e aos poucos é "esfriado"para que possa estabilizar-se num ponto mínimo. Esse processo se denomina *simulated annealing*.

2.3 Modularidade por janela

O ordenamento obtido com o algoritmo de mínimo custo deve evidenciar módulos distintos. Esses módulos podem ser identificados visualmente de forma imprecisa através de uma representação gráfica da matriz de interação, na qual os sítios de valor nulo são pontos brancos e os sítios de valor um são pontos pretos. Aglomerações de pontos pretos próximos à diagonal indicam que os nós daquela região formam módulos. Essa análise visual é útil e ilustrativa mas ao mesmo tempo insuficiente para identificar módulos interativos. Uma medida matemática torna a identificação mais clara e confiável. Com essa motivação, foi definida a medida de *modularidade por janela* [3].

Denomina-se *janela* um intervalo do ordenamento: um grupo de nós está dentro de uma janela de tamanho ω centrada em i_0 se eles estiverem posicionados dentro do intervalo $[i_0-(\omega-1)/2, i_0+(\omega-1)/2]$ no ordenamento. A *modularidade por janela* de tamanho ω do nó i_0, $mod_\omega(i_0)$, é definida pela fração de ligações compartilhadas pelos nós dentro da janela centrada em i_0 e o número total de ligações dos mesmos. Esse valor varia de zero a um: é nulo quando nenhum nó dentro da janela está ligado a outro da mesma janela e é máximo quando todas as ligações dos nós da janela são entre eles mesmos. A definição fica

$$mod_\omega(i_0) = \frac{1}{\sum_{j=i_0-\frac{\omega-1}{2}}^{i_0+\frac{\omega-1}{2}} k(j)} \sum_{i=i_0-\frac{\omega-1}{2}}^{i_0+\frac{\omega-1}{2}} k_{in}(i), \tag{2.6}$$

onde $k_{in}(i)$ é o número de ligações do nó i com nós da janela.

É útil comparar outras características da rede, como conectividade e clusterização, com a modularidade. Para tanto, é utilizada a medida *por janela* (ou *janelada*), que é o valor médio da medida dos nós dentro da janela. Para uma medida qualquer de um nó i, $f(i)$, a medida por janela é definida como

$$f_\omega(i_0) = \frac{1}{\omega+1} \sum_{i=i_0-\frac{\omega-1}{2}}^{i_0+\frac{\omega-1}{2}} f(i), \tag{2.7}$$

onde no caso da conectividade, $f(i) = k(i)$ e no caso da clusterização, $f(i) = c(i)$. Para qualquer medida janelada, inclusive a modularidade, condições periódicas de contorno são usadas.

2.4 Atribuição de funções celulares

Após usar a modularidade por janela para separar a rede em conjuntos, identifica-se as funções celulares às quais os nós de cada conjunto estão associados. Uma ferramenta muito útil para isso é o *DAVID: Functional Annotation* [9], parte de uma base de dados

disponibilizada online pelo Instituto Nacional de Saúde dos Estados Unidos (*NIH - National Institute of Health*). Submete-se uma lista de proteínas e é devolvido uma lista de termos da *Gene Ontology*[10]. A Gene Ontology (GO) é uma base de dados que descreve e classifica genes e seus respectivos produtos. Para tanto, ela reúne informações de várias outras bases de dados, para mais de 50 espécies. São três classificações, ou ontologias: componente celular (depende do local onde os produtos dos genes atuam, e.g., no núcleo), função molecular (depende da atividade que os produtos exercem, e.g., transporte de substâncias), e processo biológico (uma série de eventos moleculares pertinentes ao funcionamento do organismo, e.g., divisão celular). Portanto, insere-se uma lista de genes e é devolvida uma lista com os locais onde os produtos desses genes atuam, as funções moleculares que eles realizam e o processos biológicos dos quais eles participam. Para cada item, também é informado: o número total e os nomes dos genes da lista submetida que estão associadas àquele item; a razão entre esse número e o número total de genes da lista; e alguns valores estatísticos que informam a confiança que o item é representativo daquele grupo de genes (valor-p, Bonferroni, e Benjamini).

No DAVID, submetemos listas com os genes correspondentes às proteínas de cada módulo. Selecionamos a opção para apenas informar termos de processo biológico (GO: Biological Processes). Ressalta-se a diferença entre uma função molecular e um processo biológico: um processo biológico tem início e fim bem-definidos, e pode ocorrer em vários locais diferentes da célula (componentes celulares) e envolver várias funções moleculares. Em seguida, usamos as informações devolvidas pelo DAVID para escolher os processos mais adequados e representativos de cada módulo da rede.

Tendo escolhidos alguns termos da GO: Biological Processes para cada módulo, é possível visualizar a distribuição de proteínas participativas de cada processo dentro de toda a rede protéica. Isso é fundamental para verificar se o processo biológico corresponde principalmente a um único módulo ou se ele está distribuído entre vários. Muitas funções identificadas pelo DAVID são processos gerais como *processo metabólico celular*, dos quais a maioria das proteínas fazem parte. Para medir a distribuição da participação no processo biológico, é utilizado a medida por janela (equação (2.7)). Nesse caso, $f(i) = 1$ se a proteína correspondente ao nó i está classificada dentro do processo e $f(i) = 0$ se ela não estiver. Os dados para fazer essa análise foram obtidos usando a ferramenta AmiGO da base de dados *Gene Ontology*, também disponível online [11].

2.5 Análise de expressão gênica

Dados de expressão gênica foram obtidos da base de dados *Gene Expression Omnibus*, do Centro Nacional de Informação de Biotecnologia dos Estados Unidos *NCBI – National Center for Biotechnology Information*. Nessa base de dados estão armazenados e disponíveis dados de expressão gênica de muitos experimentos realizados por diferentes

instituições de pesquisa. As planilhas disponibilizadas contém o nome do gene e seu nível de expressão, entre outras informações que não foram relevantes para este trabalho. Esses dados foram usados para calcular a expressão por janela do ordenamento — equação (2.7) com $f(i) = \epsilon(i)$, onde ϵ é o nível de expressão do gene que codifica a proteína correspondente ao nó i. A expressão, ou atividade transcricional, por janela ao longo do ordenamento é o *transcriptograma* [3].

2.6 Diferenças entre nomenclaturas

Em todas as etapas do trabalho que envolvem dados de proteínas e genes obtidos das bases de dados referidas, foi necessário realizar traduções entre as diferentes nomenclaturas protéicas e gênicas utilizadas pela comunidade científica. A nomenclatura dos dados usados para montar a rede protéica é a *Ensembl Protein*, que não é reconhecida pelo *DAVID Functional Annotation*. Além disso, os dados de expressão gênica se referem aos genes e não às proteínas, que obviamente têm nomes diferentes. A tradução entre essas três nomenclaturas foi feita usando "dicionários" que foram obtidos do *Biomart*, um projeto do Instituto de Pesquisa em Câncer de Ontario (*OICR – Ontario Institute for Cancer Research*) e do Instituto Europeu de Bioinformática (*EBI – European Bioinformatics Institute*) [12]. Para usar o DAVID, os termos do *Ensembl Protein* foram traduzidos para *Ensembl Gene*; para os dados de expressão, foram traduzidos para *HGNC – Human Genome Nomenclature*.

Capítulo 3

Resultados

A rede protéica do *Homo sapiens* formada com os dados obtidos conforme descrito na seção 2.1 possui $N = 9019$ nós e $k_{\text{total}} = 111602$ ligações. O algoritmo de ordenamento foi aplicado com duração total de 3000 passos de Monte Carlo (um passo de Monte Carlo corresponde a N tentativas de troca). Os parâmetros utilizados foram $T_0 = 0,07H_0$, $\tau = 100$ e $\gamma = 0,4$. A figura 3.2 mostra a evolução do custo em função do tempo (passo de Monte Carlo).

3.1 Ordenamento final

A matriz de interação do ordenamento final está representada na figura 3.1. Para comparação, ao seu lado está a matriz de interação do ordenamento inicial, aleatório. A maioria dos sítios nos quais há ligação foi aproximada à diagonal, conforme o objetivo do algoritmo. Isso também pode ser verificado na figura 3.3, que mostra o decaimento da densidade de diagonal (ver equação (2.3)) do ordenamento final em contraste com a distribuição constante do ordenamento inicial.

A modularidade do ordenamento aleatório é baixa e constante. Não há como discriminar módulos nesse ordenamento. A clusterização e conectividade por janela também são relativamente constantes. Por sua vez, o ordenamento final apresenta picos e vales na modularidade, delimitando regiões de maior interatividade. O perfil da clusterização e da conectividade também muda: grupos de proteínas de maior ou menor conectividade e clusterização são aglomerados. A comparação entre essas medidas do ordenamento inicial e final está na figura 3.4, que apresenta os resultados para uma janela de $\omega = 251$.

3.2 Módulos

O tamanho dos módulos identificáveis no ordenamento depende do tamanho da janela usada para calcular a modularidade. Janelas grandes podem agrupar vários módulos

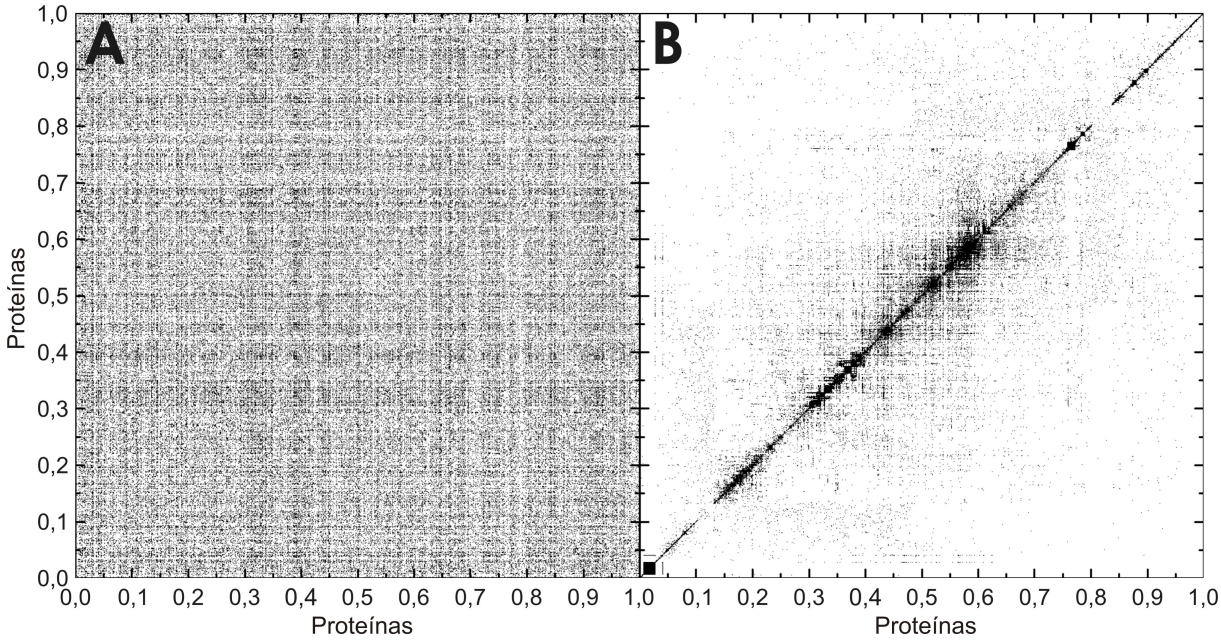

Figura 3.1: Representação da matriz de interação $M_{i,j}$ do proteoma humano no ordenamento inicial (**A**) e após 3000 passos de Monte Carlo (**B**). Espaços em branco representam sítios onde não há ligação e pontos pretos representam sítios onde há ligação. Os eixos foram normalizados pelo tamanho da rede $N = 9019$.

num único cluster, enquanto janelas menores podem dividir um grande módulo em vários subgrupos. A melhor janela a ser usada para analisar o proteoma depende da resolução desejada.

Usando a modularidade de janela $\omega = 251$, o ordenamento foi decomposto em dezenove regiões a serem analisadas separadamente (figura 3.5). As figuras 3.7 e 3.8 mostram os grafos das redes formadas pelos nós de cada região, nos quais fica claro seu caráter modular. Em função do tamanho de janela escolhido, alguns módulos podem ser decompostos em vários. Isso é facilmente observado nos grafos dos módulos 4 e 5 (os módulos compostos por todos os nós dentro da região 4 e da região 5, respectivamente). Cada um é composto principalmente por dois aglomerados; porém, a modularidade de janela $\omega = 251$ não mostra essa subestrutura claramente. A modularidade de janela $\omega = 101$ (figura 3.6), por outro lado, tem dois picos na região 4 (aproximadamente $0,22 < x < 0,25$) e dois na região 5 ($0,25 < x < 0,29$).

Outros grafos interessantes são os da região 1 e 19. A primeira é uma região de altíssima modularidade e seu grafo mostra um conjunto de nós muito conectados rodeado por nós de baixa conectividade (tipicamente pares ou trios isolados). A modularidade é muito alta porque a grande maioria dos nós pertence a este conjunto central e quase todas suas ligações são entre si. Os outros, comparativamente poucos, têm poucas ligações que os conectam a um ou outro nó da mesma região. Esse tipo de nó é minoria na região 1 e maioria na região 19. Nos dois casos, quase não ha ligação com nós fora do região, o que

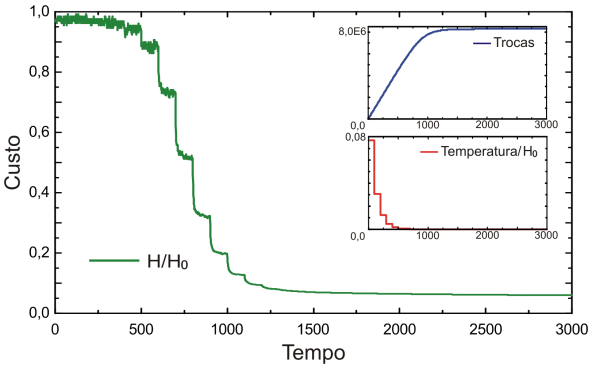

Figura 3.2: Custo em função do tempo (passos de Monte Carlo). Na direita superior, o número de trocas (em cima) e a temperatura (embaixo) em função do tempo. O custo e a temperatura foram normalizados pelo custo do estado inicial (H/H_0, T/H_0).

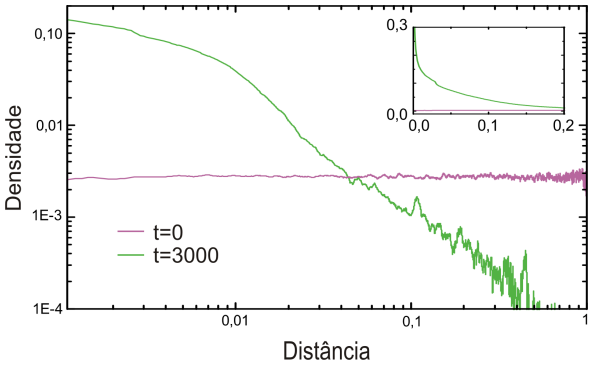

Figura 3.3: $P(M_{i,j}||i-j|=l)$ (média dentro de janela $\omega = 50$) em função de l dos ordenamentos inicial e final. O eixo horizontal l foi normalizado por $N-1$, que é a maior distância entre dois nós.

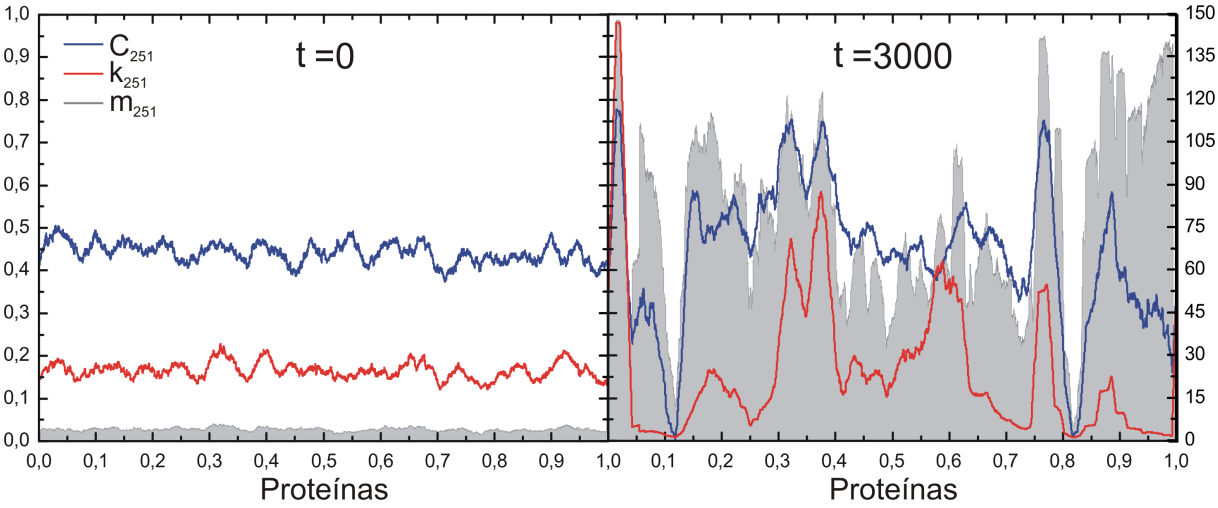

Figura 3.4: Modularidade, conectividade e clusterização por janela ($\omega = 251$) do ordenamento inicial ($t = 0$) e do ordenamento final ($t = 3000$). O eixo vertical do lado esquerdo é a escala da modularidade e da clusterização, que variam entre 0 e 1, enquanto o do lado direito é a escala da conectividade.

torna a modularidade alta.

3.3 Associação de processos biológicos

Para cada um dos dezenove módulos identificados, foi inserida a lista de proteínas no DAVID e escolhidos os processos biológicos mais relevantes. De um total de 100 processos testados, 13 foram identificados como bons representantes dos módulos (tabela 3.1). Os demais não foram incluidos ou por não estarem suficientemente presentes, ou por não estarem concentrados num único módulo, ou por já serem contemplados por um dos processos já escolhidos (por exemplo, processamento de mRNA está incluido em processamento de

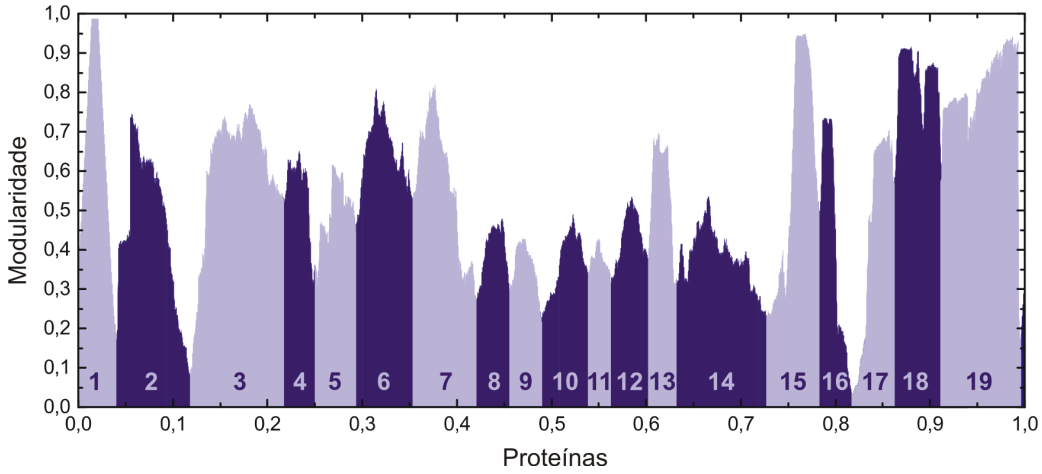

Figura 3.5: Numeração dos módulos identificados ($\omega = 251$).

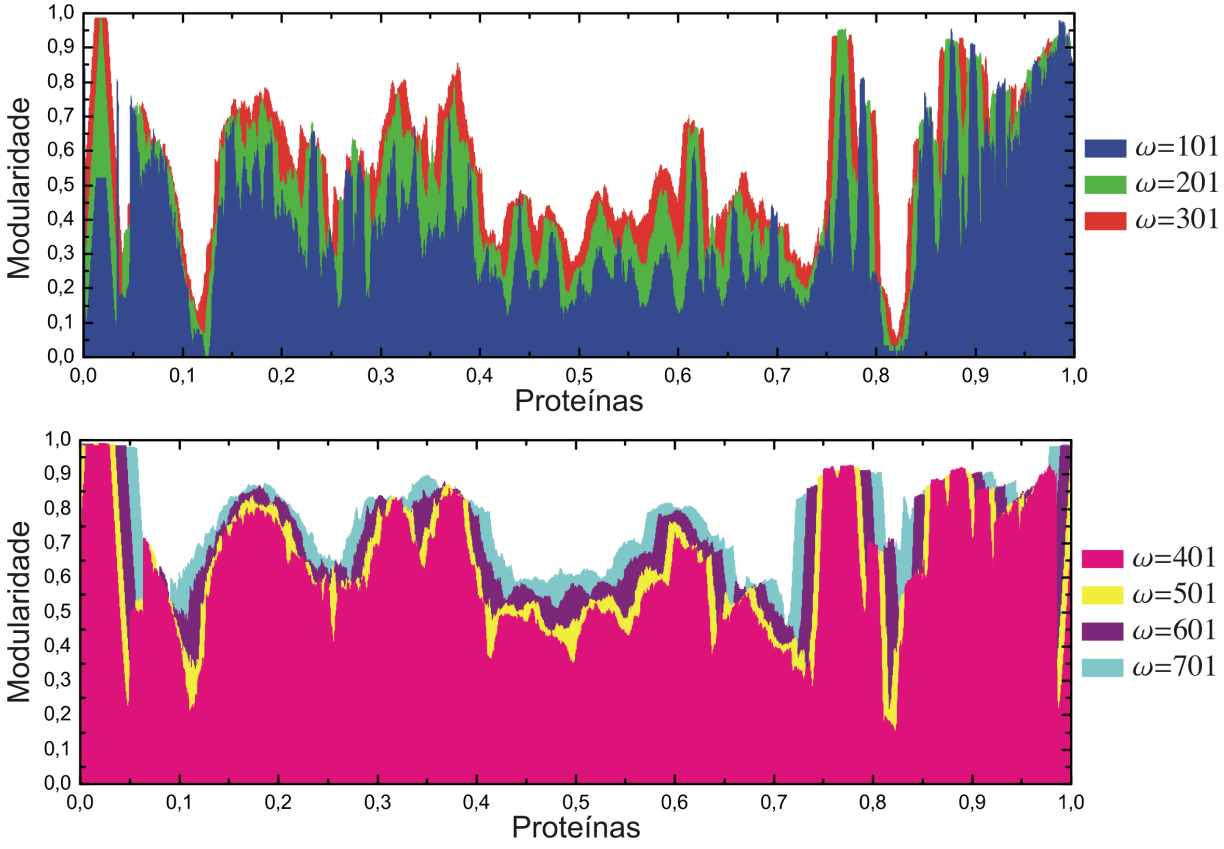

Figura 3.6: Modularidade para diferentes janelas.

RNA). A figura 3.9 mostra a distribuição das funções escolhidas no ordenamento.

3.4 Transcriptograma

Perfis de expressão gênica de pacientes em diferentes fases de câncer de pulmão foram obtidos do GEO da série GSE10072 (ver seção 2.5). Os perfis foram realizados usando

Módulo	Processo biológico	Código GO	%	Definição
1	Percepção sensorial de estímulo químico	GO:0007606	$47,04\%$	A série de eventos necessários para um organismo receber um estímulo químico sensorial, convertê-lo em um sinal molecular, e reconhecer e caracterizar o sinal.
3	Processo metabólico de ácido orgânico	GO:0006082	$31,96\%$	As reações químicas e rotas envolvendo ácidos orgânicos.
6	Processo de ciclo celular	GO:0022402	$32,17\%$	Processo celular envolvido na progressão de fases bioquímicas e morfológicas que ocorrem numa célula durante eventos sucessivos de replicação celular ou nuclear.
7	Processamento de RNA	GO:0006396	$21,22\%$	Processo envolvido na conversão de um ou mais transcritos primários de RNA em um ou mais moléculas maduras de RNA.
8	Regulação negativa de expressão gênica	GO:0010629	$26,60\%$	Processo que diminui a frequência, taxa ou extensão de expressão gênica.
9	Rota de sinalização da proteína transmembrana receptora serina/treonina	GO:0007178	$20,16\%$	Série de sinais moleculares gerados pela ligação da proteína transmembrana receptora serina/treonina.
10	Apoptose	GO:0006915	$27,03\%$	Tipo de morte celular programada.
12	Adesão celular	GO:0007155	$28,70\%$	A ligação de uma célula a outra célula ou a um substrato através de moléculas de adesão celular.
13	Transdução de sinal mediado por GTPase	GO:0007264	$20,18\%$	Série de sinais moleculares transmitidos por GTPase monomérica pequena.
15	Tradução	GO:0006412	$26,97\%$	Processo metabólico pelo qual uma proteína é formada.

Tabela 3.1: As principais funções de cada módulo. A segunda coluna informa o código que identifica a função no Gene Ontology. A terceira é a razão entre o número de proteínas do módulo que fazem parte do processo e o tamanho do módulo. A última é a definição, dada no Gene Ontology, do processo biológico (algumas foram resumidas).

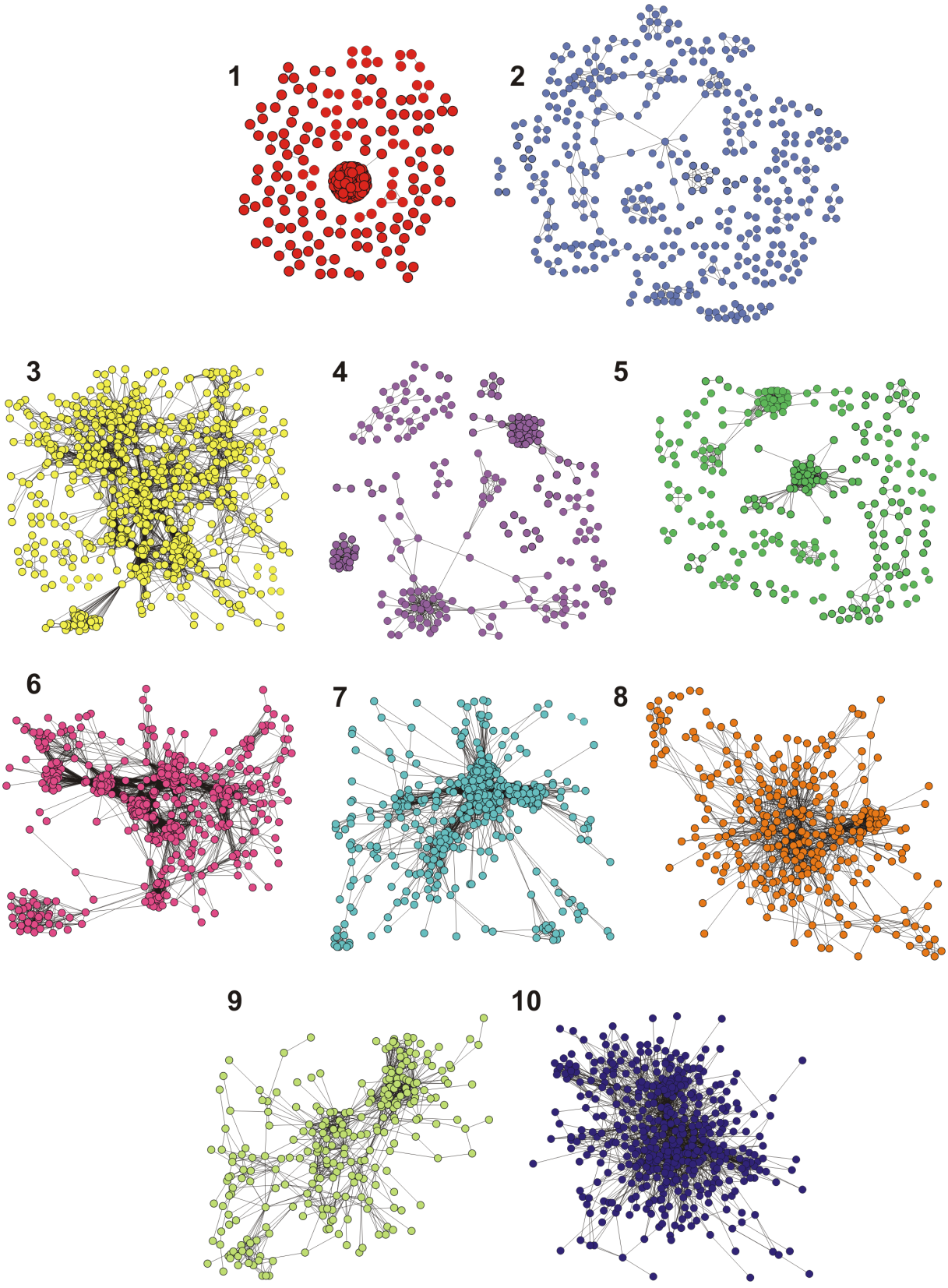

Figura 3.7: Grafos dos módulos das regiões 1 a 10, feitos usando a ferramenta *Medusa*[13].

microarrays de RNA e fazem parte de um estudo sobre as alterações moleculares causadas

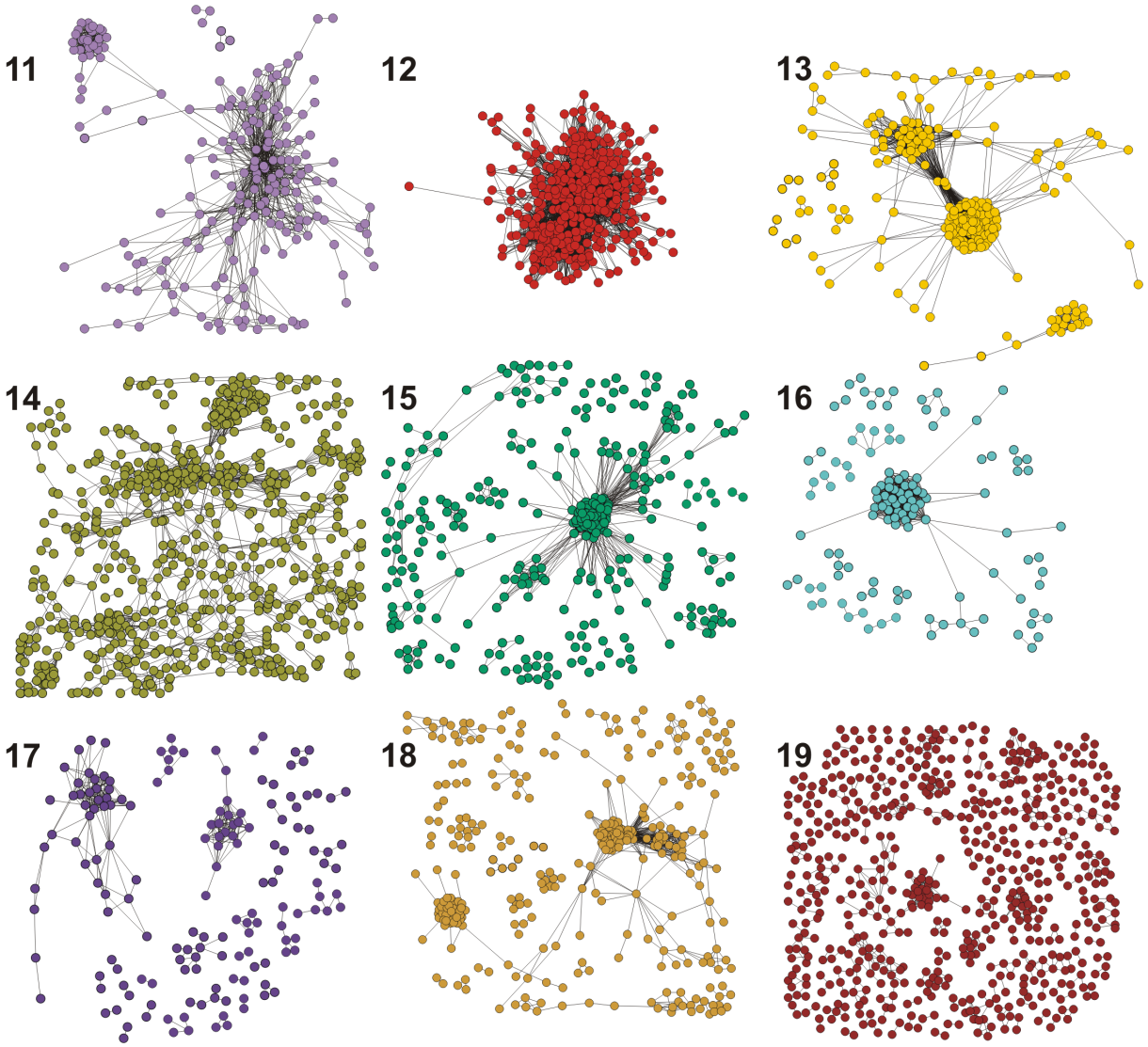

Figura 3.8: Grafos dos módulos das regiões 11 a 19, feitos usando a ferramenta *Medusa*[13].

pelo fumo de tabaco e sua relação com câncer pulmonar [14]. Usamos os dados de 33 pacientes: 11 não-fumantes (nunca fumaram), 10 ex-fumantes e 12 fumantes (tabela 3.2). Para cada um, há amostras de tecido pulmonar com adenocarcinoma e de tecido pulmonar normal.

Os níveis de expressão por janela ($\epsilon_\omega(i)$) foram calculados para cada amostra . Em seguida, foi calculada o valor médio de $\epsilon_\omega(i)$ das amostras de tecido normal dos não-fumantes: $\epsilon^0_\omega(i)$. Esse valor serve como base de comparação para os níveis de expressão das outras amostras; por isso, todas foram normalizadas por ele (inclusive cada amostra normal de não-fumante, individualmente). Para essa comparação, também foi calculado o desvio-padrão amostral dos tecidos normais de não-fumantes,

$$s_\omega(i) = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(\frac{\epsilon_\omega(i)}{\epsilon^0_\omega(i)} - 1\right)^2}. \tag{3.1}$$

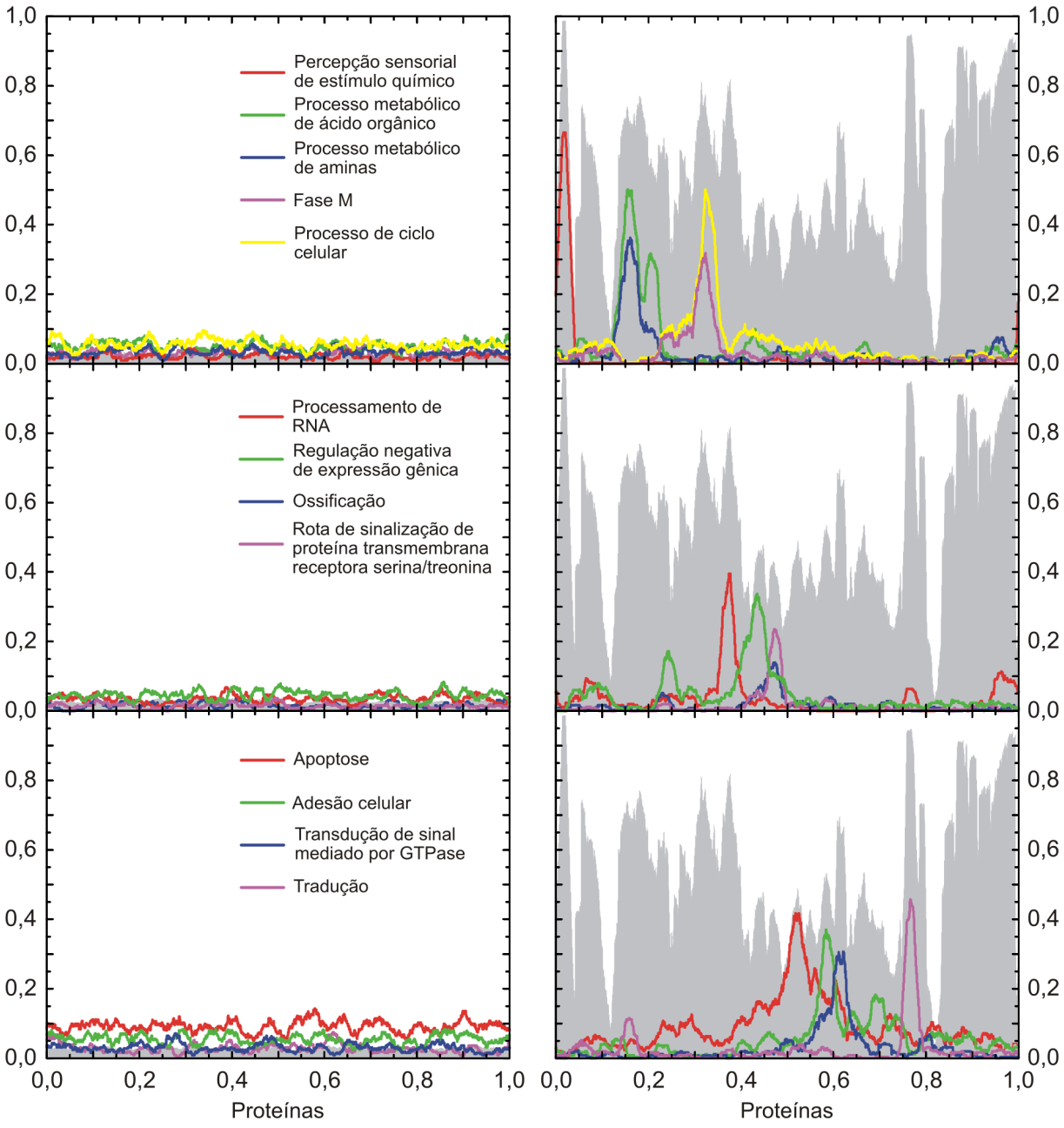

Figura 3.9: Distribuição dos processos biológicos no ordenamento inicial (esquerda) e final (direita). Ao fundo, a modularidade ($\omega = 251$).

A figura 3.10 mostra os transcriptogramas de cada amostra para a janela $\omega = 251$. Os valores fora da região rosa ou amarela mostram uma variação de mais de duas ou quatro vezes $s_\omega(i)$, respectivamente, em relação à média dos tecidos normais de não fumantes. As figuras 3.11, 3.12 e 3.13 mostram cada transcriptograma individual das amostras de fumantes, ex-fumantes e não-fumantes, respectivamente.

Os transcriptogramas mostram variação acentuada nas amostras de fumantes e ex-fumantes em relação aos não-fumantes, tanto no tecido normal quanto no canceroso. Os valores médios de atividade transcricional por janela de cada grupo mostra maior variação

Não-Fumante				Ex-Fumante				Fumante			
Amostra	Sexo	Idade	Estádio	Amostra	Sexo	Idade	Estádio	Amostra	Sexo	Idade	Estádio
GT00006	M	69	IIB	GT00059	M	67	IIIA	GT01003	F	45	IV
GT01001	F	70	IIIB	GT00146	M	71	IIB	GT01025	M	65	IA
GT01011	F	67	IIB	GT01099	M	71	IB	GT01038	F	46	IIIA
GT01036	F	67	IB	GT01100	M	69	IIB	GT01097	M	59	IA
GT01061	F	68	IIIA	GT01194	M	67	IB	GT01107	M	59	IA
GT01105	M	61	IV	GT01232	M	81	IB	GT01117	M	63	IB
GT01119	F	74	IIIA	GT01243	M	73	IA	GT01120	M	51	IIB
GT01130	F	71	IB	GT01246	M	76	IIB	GT01121	M	52	IV
GT01182	F	78	IB	GT01445	M	70	IIB	GT01128	M	66	IB
GT01247	F	62	IIIA					GT01129	F	55	IB
GT01421	M	68	IB					GT01222	M	63	IB
								GT01233	M	57	IIB

Tabela 3.2: Informações sobre amostras.

no tecido canceroso de fumantes, principalmente nas regiões dos módulos 3 e 6, nos quais a transcrição é mais intensa, e 13 e 14, onde é menos intensa (figura 3.14).

No tecido normal, as médias dos fumantes e ex-fumantes flutuam muito menos que no tecido cancerosos. Por outro lado, o desvio médio das amostras em torno de 1 mostra que o perfil de transcrição nesses casos está alterado em relação ao tecido normal dos não-fumantes (figura 3.15).

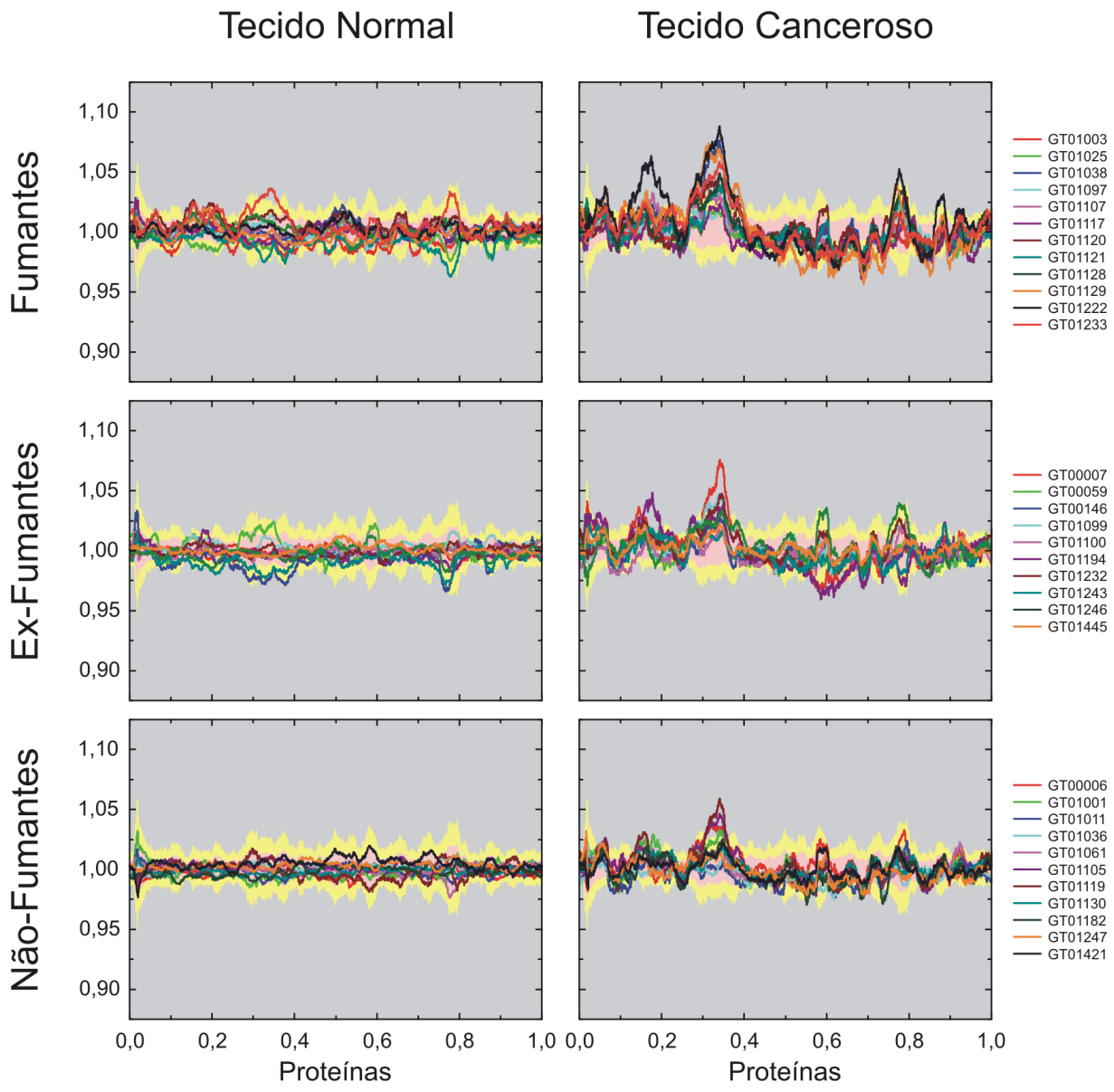

Figura 3.10: Transcriptograma das amostras de tecido canceroso e normal dos fumantes, ex-fumantes e não-fumantes, para a janela $\omega = 251$ ($\frac{\epsilon_{251}(i)}{\epsilon^0_{251}(i)}$ em função de i). Ao fundo, o intervalo de cor rosa representa $[1 - 2s_{251}(i), 1 + 2s_{251}(i)]$ e o de cor amarelo é $[1 - 4s_{251}(i), 1 + 4s_{251}(i)]$, onde $s_{251}(i)$ é o desvio-padrão amostral de $\frac{\epsilon_{251}(i)}{\epsilon^0_{251}(i)}$.

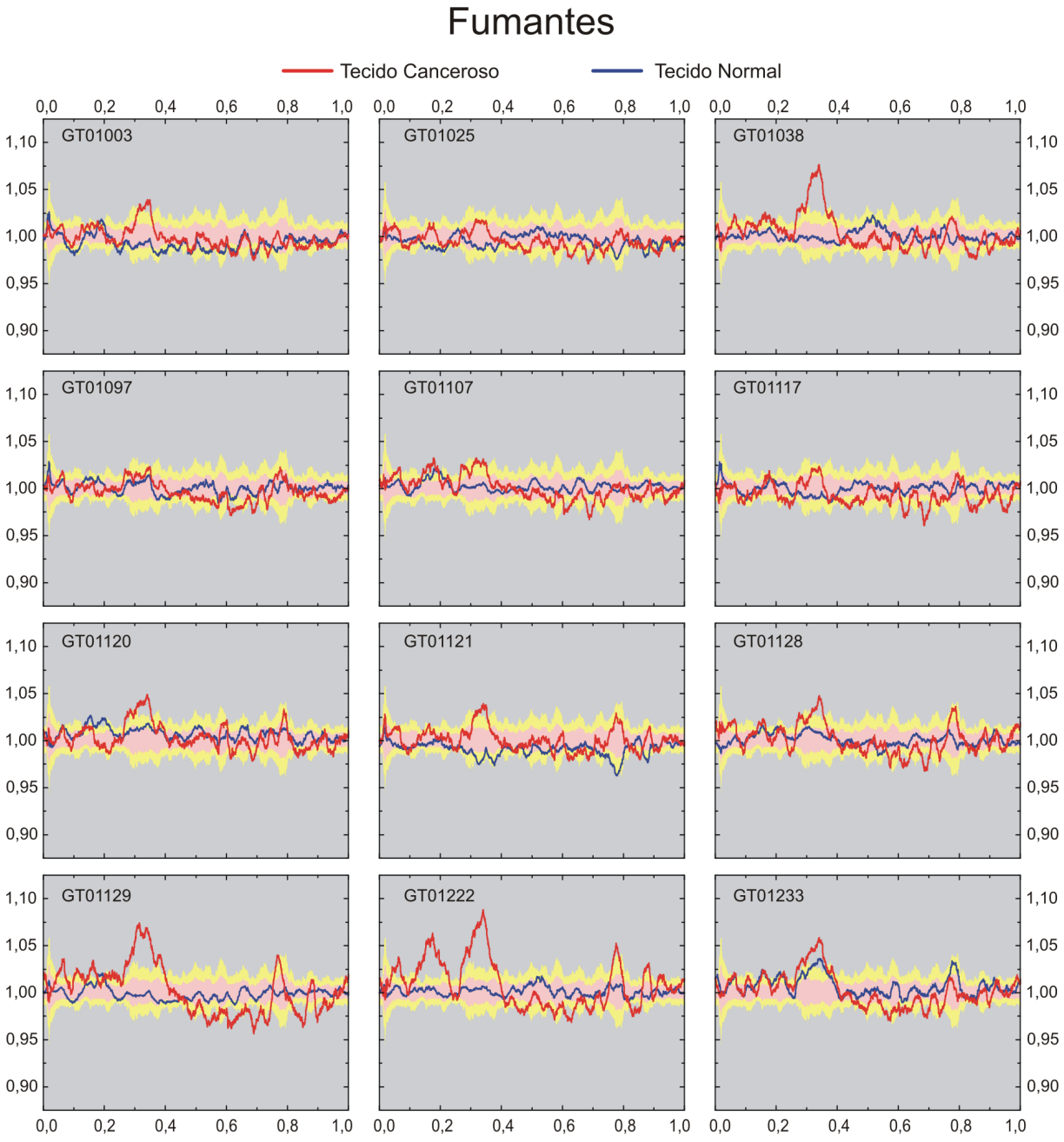

Figura 3.11: Transcriptograma das amostras de tecido canceroso e normal dos fumantes para a janela $\omega = 251$ ($\frac{\epsilon_{251}(i)}{\epsilon^0_{251}(i)}$ em função de i). Ao fundo, o intervalo de cor rosa representa $[1 - 2s_{251}(i), 1 + 2s_{251}(i)]$ e o de cor amarelo é $[1 - 4s_{251}(i), 1 + 4s_{251}(i)]$, onde $s_{251}(i)$ é o desvio-padrão amostral de $\frac{\epsilon_{251}(i)}{\epsilon^0_{251}(i)}$.

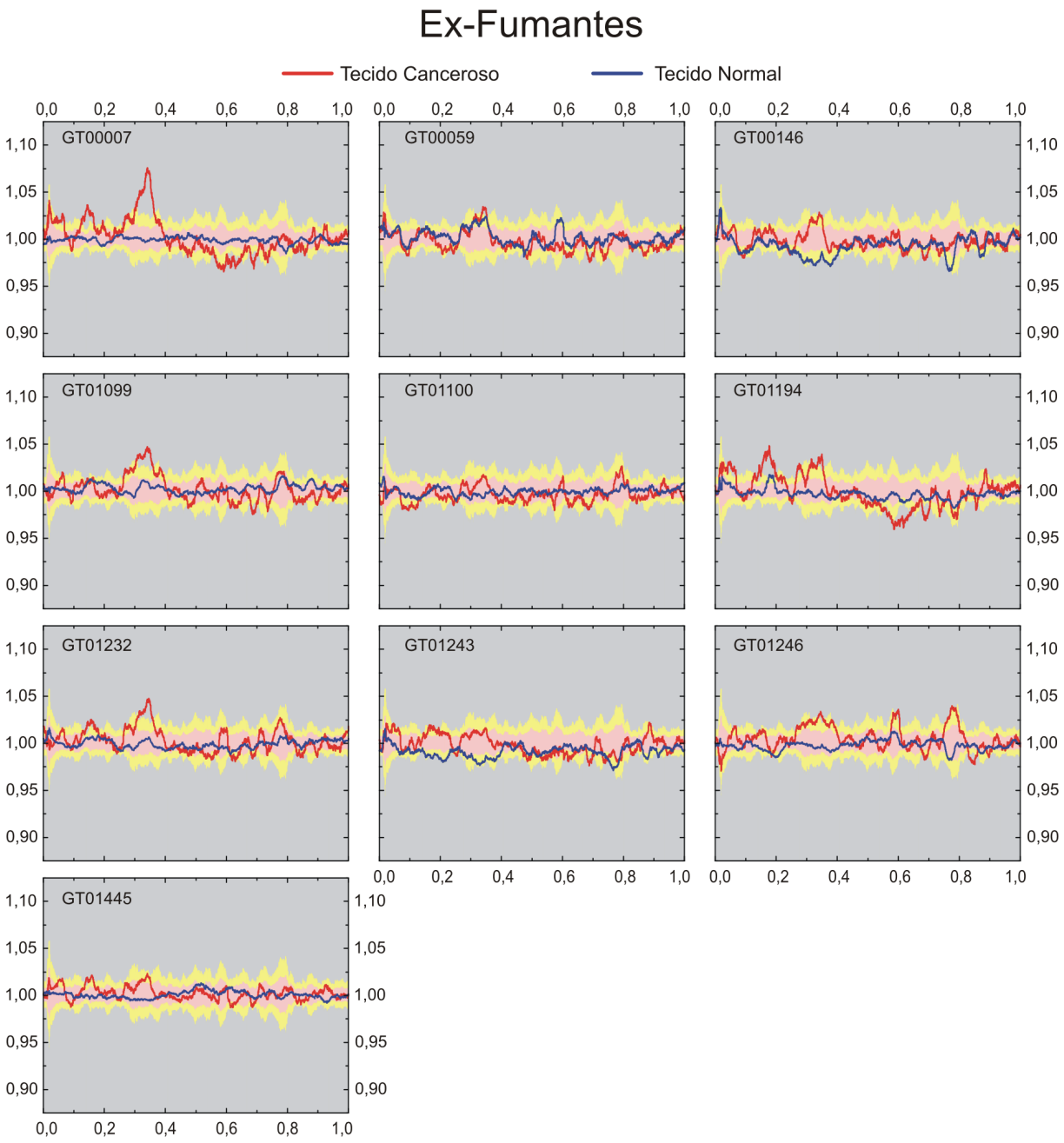

Figura 3.12: Transcriptograma das amostras de tecido canceroso e normal dos ex-fumantes para a janela $\omega = 251$ ($\frac{\epsilon_{251}(i)}{\epsilon^0_{251}(i)}$ em função de i). Ao fundo, o intervalo de cor rosa representa $[1-2s_{251}(i), 1+2s_{251}(i)]$ e o de cor amarelo é $[1-4s_{251}(i), 1+4s_{251}(i)]$, onde $s_{251}(i)$ é o desvio-padrão amostral de $\frac{\epsilon_{251}(i)}{\epsilon^0_{251}(i)}$.

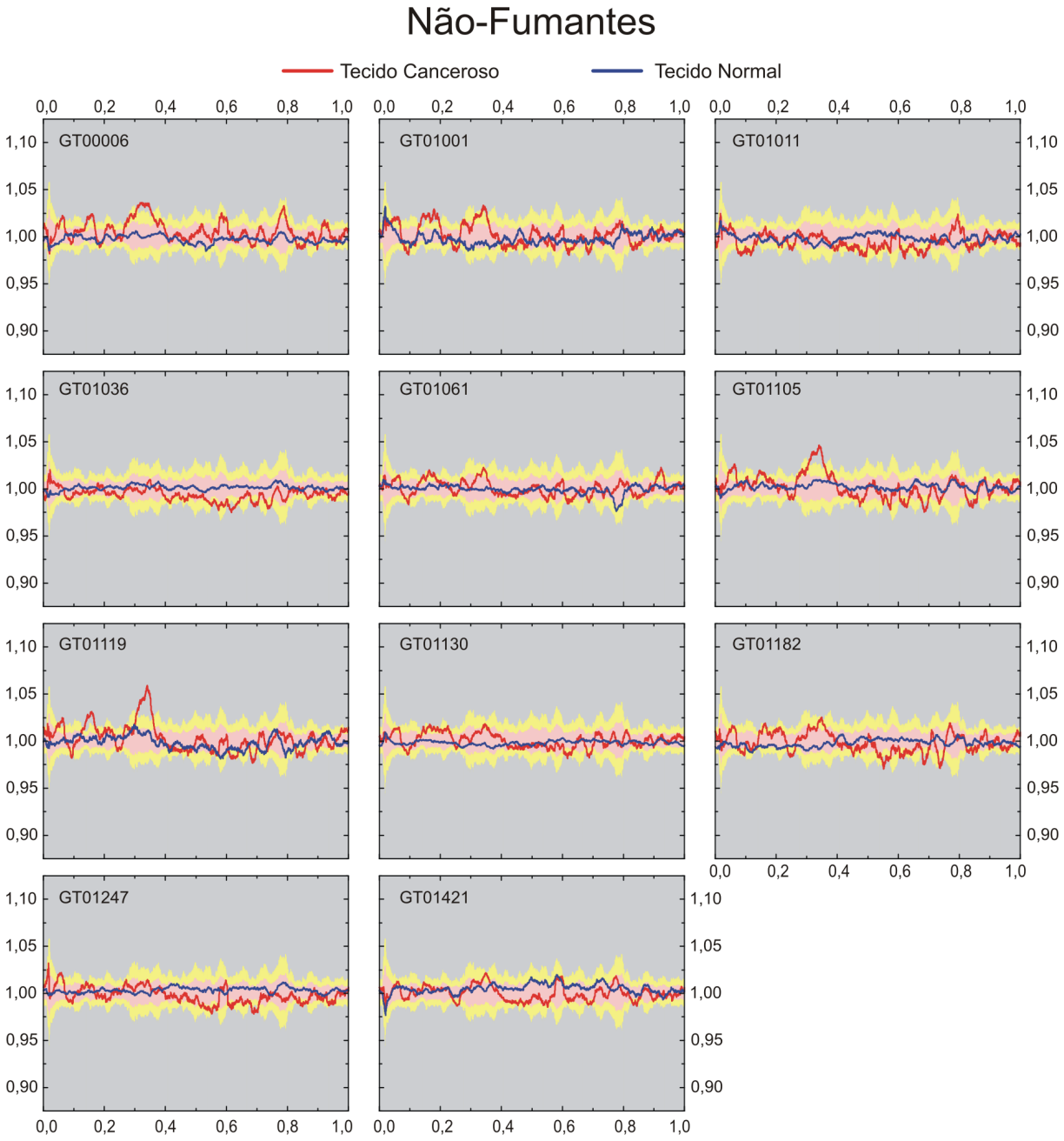

Figura 3.13: Transcriptograma das amostras de tecido canceroso e normal dos não-fumantes para a janela $\omega = 251$ ($\frac{\epsilon_{251}(i)}{\epsilon^0_{251}(i)}$ em função de i). Ao fundo, o intervalo de cor rosa representa $[1-2s_{251}(i), 1+2s_{251}(i)]$ e o de cor amarelo é $[1-4s_{251}(i), 1+4s_{251}(i)]$, onde $s_{251}(i)$ é o desvio-padrão amostral de $\frac{\epsilon_{251}(i)}{\epsilon^0_{251}(i)}$.

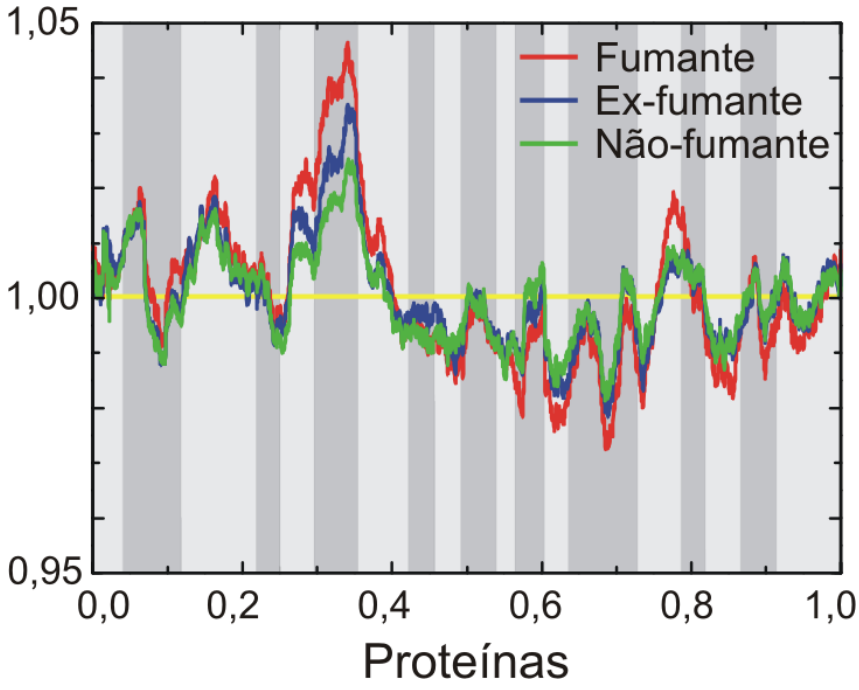

Figura 3.14: Transcrição média por janela das amostras de tecido canceroso de fumantes, ex-fumantes e não-fumantes. As listras do fundo representam as 19 regiões modulares e foram colocadas para facilitar a identificação dos módulos nos quais a expressão gênica está alterada.

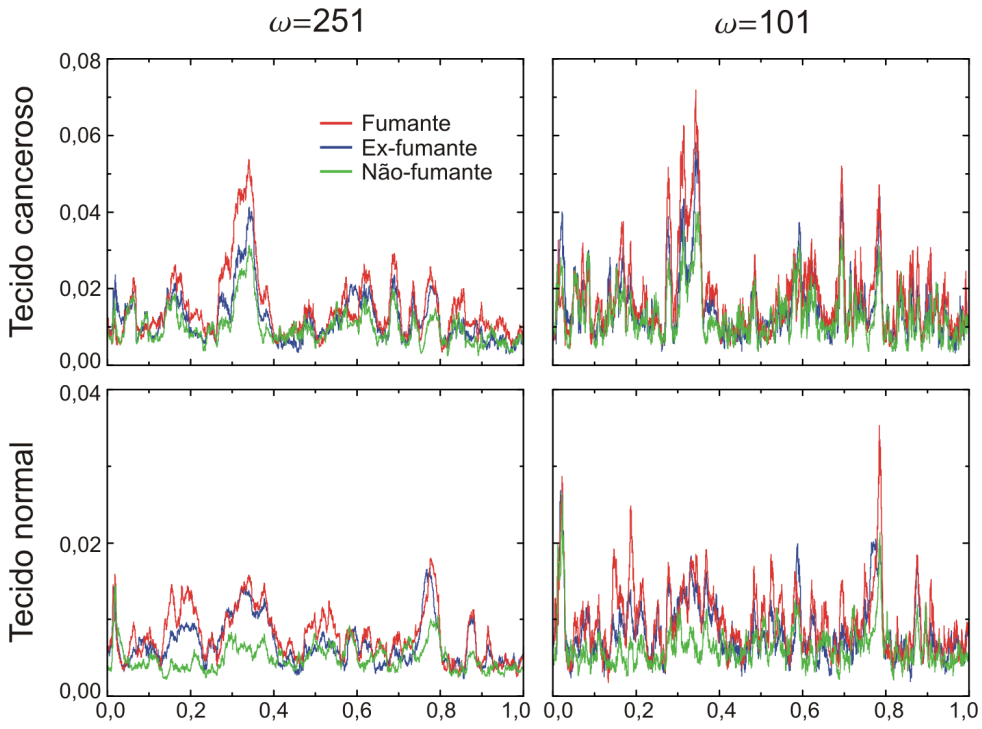

Figura 3.15: Desvios das amostras em torno de 1 (ver equação (3.1)) de tecido canceroso e normal para $\omega = 251$ e $\omega = 101$.

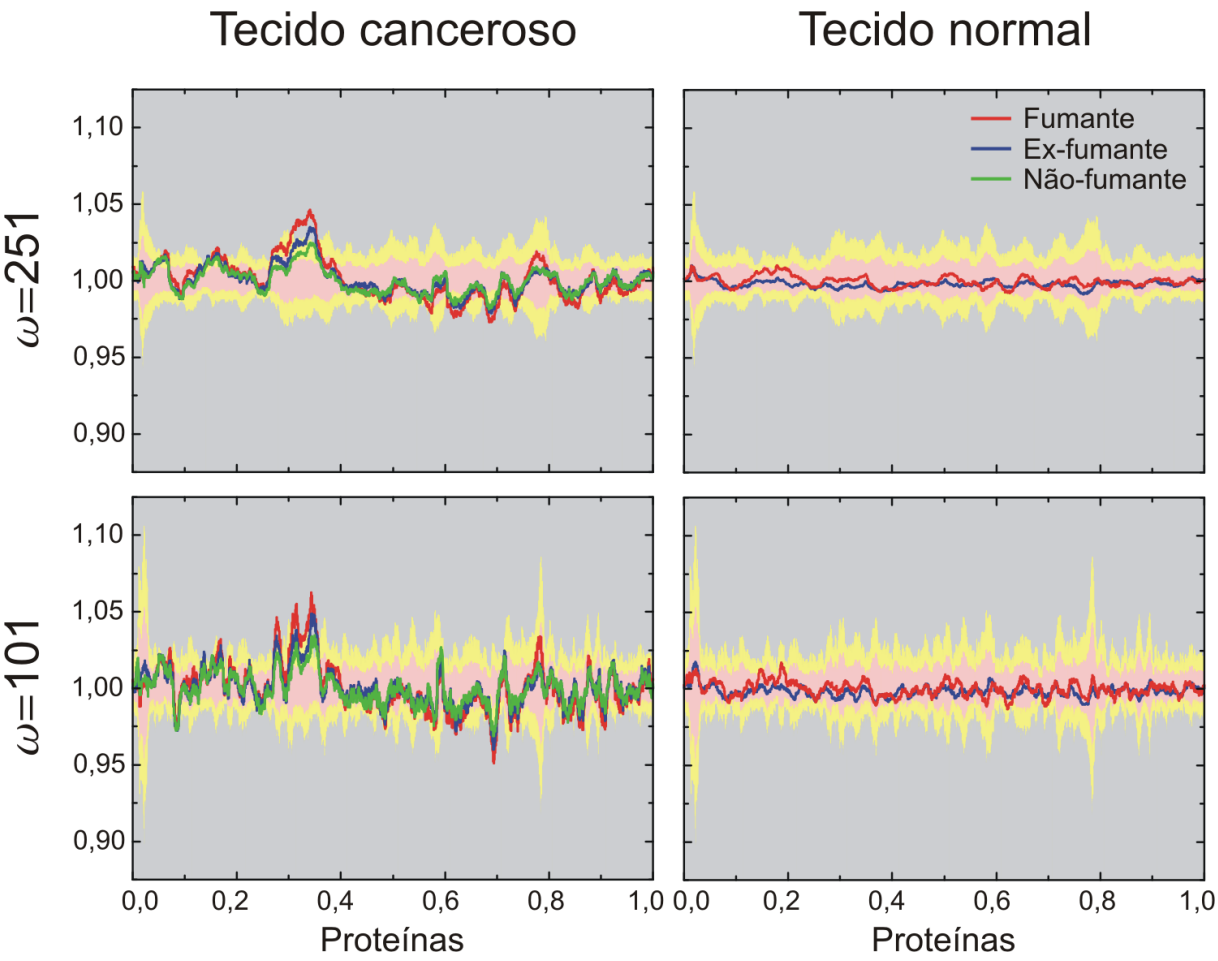

Figura 3.16: Transcrição média por janela ($\omega = 101$ e $\omega = 251$) das amostras de tecido canceroso de fumantes, ex-fumantes e não-fumantes, e de tecido normal de fumantes e ex-fumantes (no caso do tecido normal de não-fumantes, seria simplesmente 1, já que todos foram normalizados por essa média).

Capítulo 4

Conclusões

Conforme declarado na seção 1.1, os três objetivos principais desse trabalho eram:

1. Organizar a rede protéica do *Homo sapiens* de forma a evidenciar módulos distintos.
2. Identificar processos biológicos que sejam representativos de cada módulo.
3. Usar o transcriptograma para analisar e comparar a atividade transcricional de células cancerosas e células saudáveis de fumantes, ex-fumantes e não-fumantes diagnosticados com câncer pulmonar.

Cada um foi cumprido de acordo com os métodos propostos:

1. Ordenamento da rede através do algoritmo de minimização da função custo (seções 2.1 e 2.2).
2. Decomposição da rede em módulos usando a modularidade por janela (seção 2.3) e cálculo da distribuição dos processos biológicos no ordenamento (seção 2.4).
3. Cálculo da atividade transcricional por janela de cada amostra (seção 2.5).

O capítulo 3 mostra 19 módulos identificados usando o algoritmo proposto, confirmando a estrutura modular da rede e demonstrando a eficácia do método para decompor a rede em subsistemas menores. Dos 19 módulos, dez são compostos por proteínas envolvidas nas mesmas funções celulares (seção 3.3), e isso representa um resultado ainda mais importante do algoritmo: os módulos identificados não são apenas grupos de proteínas que interagem entre si — eles também representam processos biológicos.

Finalmente, o último e principal objetivo (aquele que dá o título a esse trabalho), também foi atingido. Os transcriptogramas expostos na seção 3.4 mostram claramente diferenças significativas de atividade transcricional entre os tipos de amostra de tecido: células de tecido canceroso têm regiões muito alteradas em relação às células de tecido normal; células de fumantes são um pouco mais alteradas que as de ex-fumantes, e ambas

são mais alteradas que as de não-fumantes (tanto o tecido canceroso quanto o tecido normal).

Além disso, os transcriptogramas mostram que há regiões onde as alterações são mais acentuadas. Há atividade transcricional mais intensa nos módulos 3 e 6: a média dos tecidos cancerosos têm uma diferença de mais de 4 vezes o desvio-padrão amostral das normais não-fumantes nessas regiões. Como o módulo 6 está associado a processos de ciclo celular (tabela 3.1 e figura 3.9), conclui-se que células de tecido pulmonar de adenocarcinoma, em média, apresentam um perfil de expressão gênica com maior produção de proteínas ligadas ao ciclo celular. Esse resultado está em concordância com o resultado do estudo original de Landi et al [14], no qual os autores identificam transcrição mais intensa nos genes de ciclo celular.

Há regiões no transcriptograma onde a atividade transcricional do tecido canceroso, em média, chega a estar 4 vezes o desvio-padrão amostral menor que a atividade das normais não-fumantes. Duas delas são a região 13 e no centro da região 14. A distribuição da adesão celular no ordenamento possui um pico aproximadamente no meio do módulo 14 (figura 3.9). Proteínas de adesão celular, portanto, são produzidas a uma taxa menor nas células cancerosas da amostra.

O artigo de Landi et al também afirma que a transcrição nas células de tecido normal de fumantes e ex-fumantes é significativamente mais alterada que as de tecido normal de não-fumantes. Isso é mais visível olhando para o conjunto de transcriptogramas das amostras individuais (figura 3.10) do que para a média das amostras de cada tipo (figura 3.16), pois na média algumas variações se anulam. Outra maneira de confirmar essa variação é na comparação entre os desvios (figura 3.15), onde os desvios de tecido normal de fumantes e ex-fumantes não têm muita diferença, mas ambos são maiores que o desvio de tecido normal de não-fumantes.

A grande utilidade do transcriptograma é a facilidade de identificação de regiões de maior alteração transcricional e a possibilidade de comparar as regiões alteradas com os processos biológicos do módulo correspondente. É uma ferramenta para analisar o perfil global de expressão gênica de uma célula. Como mostram os resultados apresentados, esses perfis possuem padrões de acordo com o tipo e condição da célula, o que possibilita o uso do transcriptograma como ferramenta de diagnóstico.

Perspectivas

A validação do transcriptograma como ferramenta de diagnóstico precisa de mais resultados de diferentes amostras e experimentos. Isso pode ser realizado obtendo mais dados de transcrição do Gene Expression Omnibus. Um aspecto dos dados utilizados nesse trabalho que pode ser explorado é analisar a possível correlação entre alteração transcricional e o estádio do adenocarcinoma, principalmente nas regiões do ordenamento que são mais

associadas a padrões tumorais.

Um aspecto importante que precisa ser melhorado é a interpretação dos picos de modularidade e os processos biológicos associados a eles. Isso será feito com a colaboração do Departamento de Bioquímica da UFRGS. Especialistas podem indicar outros termos da GO: Biological Processes que sejam mais interessantes para nossa análise, além de interpretar mais profundamente os transcriptogramas.

Outra área desse trabalho que pode ser aprofundada é a análise modular de diferentes janelas. A modularidade de diferentes janelas foi apresentada (figura 3.6) mas apenas a janela $\omega = 251$ foi utilizada para discriminar módulos. Janelas de tamanho menor, como $\omega = 101$ e $\omega = 151$, podem discriminar processos biológicos mais específicos. Transcriptogramas de janela $\omega = 101$ foram apresentados (figuras 3.16, 3.15) para mostrar que talvez as alterações transcricionais sejam ainda mais concentradas em regiões menores, mas não foram analisados.

Há muitas possibilidades para desenvolver essa pesquisa, e esses aspectos citados são apenas os primeiros passos para aprofundar nosso conhecimento da estrutura do proteoma e aperfeiçoar o transcriptograma como ferramenta de diagnóstico.

Referências Bibliográficas

[1] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[2] Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113, 2004.

[3] José Luiz Rybarczyk Filho, Mauro A. A. Castro, Rodrigo J. S. Dalmolin, José C. F. Moreira, Leonardo G. Brunnet, and Rita M. C. de Almeida. Towards a genome-wide transcriptogram: the *Saccharomyces cerevisiae* case. *Nucleic Acids Research*, 2010. Aceito para publicação.

[4] Reihard Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, 4 edition, 2010.

[5] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[6] Albert L. Lehninger, David L. Nelson, and Michael M. Cox. *Principles of Biochemistry*. Worth, 2 edition, 1993.

[7] Lars J. Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering. String 8 - a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37:D412–D416, 2009.

[8] Christian von Mering, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathilde Foglierini, Nelly Joufre, Martijn A. Huynen, and Peer Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33:D433–D437, 2005.

[9] Glynn Dennis, Brad T. Sherman, Douglas A. Hosack, Jun Yang, Wei Gao, H. Clifford Lane, and Richard A. Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5):P3, 2003.

[10] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[11] Seth Carbon, Amelia Ireland, Christopher J. Mungall, ShengQiang Shu, Brad Marshall, and Suzanna Lewis. Amigo: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009.

[12] Syed Haider, Benoit Ballester, Damian Smedley, Junjun Zhang, Peter Rice, and Arek Kasprzyk. Biomart central portal - unified access to biological data. *Nucleic Acids Research*, 37:W23–W27, 2009.

[13] Sean D. Hooper and Peer Bork. Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, 21(24):4432–4433, 2005.

[14] Maria T. Landi, Tatiana Dracheva, Melissa Rotunno, Jonine D. Figueroa, Huaitian Liu, and et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS ONE*, 3(2):e1651, 2008.