República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) PI 1106018-2 A2

(22) Data de Depósito: 14/12/2011
(43) Data da Publicação: 19/11/2013
(RPI 2237)

(51) Int.Cl.:
G06F 19/10
G06F 19/18

(54) Título: MÉTODO, SISTEMA E APARATO DE ANÁLISE DE DADOS DE EXPRESSÃO GÊNICA (TRANSCRIPTOGRAMA).

(73) Titular(es): Universidade Federal do Rio Grande do Sul

(72) Inventor(es): Diego Bonatto, José Cláudio Fonseca Moreira, José Luiz Rybarczyk Filho, Leonardo Gregory Brunnet, Mauro Antônio Alves Castro, Rita Maria Cunha de Almeida, Rodrigo Juliani Siqueira Dalmolin

(57) Resumo: MÉTODO, SISTEMA E APARATO DE ANÁLISE DE DADOS DE EXPRESSÃO GÊNICA (TRANSCRIPTOGRAMA) A presente invenção descreve um novo e inventivo método, sistema e aparato de análise para dados de transcrição gênica. Preferencialmente a análise dos dados de transcrição é realizada através de um ordenamento da informação biológica de determinado organismo (genes ou proteínas a eles associadas) de tal maneira que genes cujas proteínas têm maior probabilidade de estarem associadas de alguma forma estão mais próximos nesta lista ordenada, permitindo a produção de um meio de visualização dessa interação (aqui também conhecido como "transcriptograma") demonstrando o nível de expressão de cada RNA mensageiro. A análise aqui apresentad inclui também um tratamento estatístico o qual evidencia alterações no metabolismo celular com precisão não atingida por outros métodos.

Relatório Descritivo de Patente de Invenção

MÉTODO, SISTEMA E APARATO DE ANÁLISE DE DADOS DE EXPRESSÃO GÊNICA (TRANSCRIPTOGRAMA)

Campo da Invenção

A presente invenção descreve um novo e inventivo método, sistema e aparato de análise para dados de transcrição gênica. Preferencialmente a análise dos dados de transcrição é realizada através de um ordenamento da informação biológica de determinado organismo (genes ou proteínas a eles associadas) de tal maneira que genes cujas proteínas têm maior probabilidade de estarem associadas de alguma forma estão mais próximos nesta lista ordenada, permitindo a produção de um meio de visualização dessa interação (aqui também conhecido como "transcriptograma") demonstrando o nível de expressão de cada RNA mensageiro. A análise aqui apresentada inclui também um tratamento estatístico o qual evidencia alterações no metabolismo celular com precisão não atingida por outros métodos. A presente invenção se situa nos campos da bioquímica, biologia molecular e bioinformática.

Antecedentes da Invenção

Dados de expressão de genoma inteiro consistem na informação a respeito do nível de expressão de milhares de genes e a análise conjunta destes dados representa um desafio. As formas usuais de abordagem comparam os níveis de expressão de células em estados modificados relativamente a um controle previamente estabelecido. Os genes são então ordenados pela co-variação de expressão em relação ao controle e aqueles genes que apresentam as variações mais significativas são selecionados para serem analisados com mais detalhe. Por outro lado, a dinâmica de expressão gênica é determinada pela ação de uma rede de genes e alterações moderadas em muitos genes interagentes podem causar efeitos mensuráveis no metabolismo celular. Estes efeitos podem não ser detectados quando se usa o critério de considerar somente os genes cujas expressões estão

maximamente alteradas, deixando de lado redes de muitos genes com expressão moderadamente alterada. No entanto, esse critério segue sendo utilizado porque reduz o número de genes a ser considerado, uma vez que um número muito grande de dados pode tornar a análise não factível.

Existem diversas técnicas para tratar de grande quantidade de dados. Um exemplo comum de filtragem de dados pode ser encontrado em processamento de fotografias digitais de alta resolução: embora o arquivo digital contenha informação sobre um número muito grande de pixels, muito maior que o número de pixels numa tela de computador, uma imagem do objeto fotografado ainda pode ser produzida na tela. Ferramentas numéricas de processamento de imagem atribuem a cada pixel da tela alguma média da informação armazenada em um grupo de pixels digitais próximos entre si, reduzindo assim a quantidade total de informação enviada para a tela, mas ainda preservando a informação global. Observe que *zooms* podem ser aplicados a essas imagens para obter imagens parciais de tal maneira que, após o *zoom*, cada pixel da tela corresponde a uma média sobre um número menor de pixels digitais. Em outras palavras, uma grande coleção de dados sobre um dado fenômeno pode ser apresentada em uma imagem mais grosseira, porém global, ou mais precisa, mas parcial. Neste exemplo, o ponto chave é a média da informação armazenada em pixels *vizinhos.* Mais ainda, a média sobre os pixels vizinhos age no sentido de neutralizar flutuações espúrias causadas por eventuais efeitos externos aleatórios.

Níveis de expressão de genes podem diferir por grandes quantidades; consequentemente uma lista aleatória de genes gera um perfil de expressão genética relativa com flutuações tão fortes que pouca, se alguma, informação pode ser extraída. Técnicas para suavizar flutuações de perfis consideram em geral médias tomadas sobre pontos vizinhos. No caso de uma lista de genes ordenados seguindo algum critério que favoreça o agrupamento de genes interagentes, a distância entre quaisquer dois genes desta lista correlacionaria com a probabilidade que seus produtos gênicos estejam associados, fornecendo um critério natural para definição de vizinhança para a lista de

genes.

Esta invenção propõe um processo para a construção de 'imagens' de expressão gênica de genomas inteiros pela produção de perfis de expressão para transcriptomas. A ideia do método é considerar médias de dados de expressão sobre genes próximos quando arranjados em uma lista. Por um lado este procedimento visa uma medida global dos dados de expressão de genoma inteiro. Por outro lado, tal processo requer a definição adequada de vizinhança entre os genes quando dispostos em uma lista, o que não é trivial.

Em uma realização preferencial, esta invenção consiste em um processo para ordenar uma lista de genes usando o método de física computacional conhecido como Monte Carlo [1]. O objetivo é produzir uma lista de genes onde aqueles mutuamente interagentes estejam próximos, de tal maneira que a distância entre dois genes da lista esteja correlacionada com a probabilidade que seus produtos gênicos estejam associados. O critério para associação ou não de duas proteínas pode ser obtido de bancos de dados públicos como o STRING [2]. Uma primeira vantagem em relação ao agrupamento por co-expressão gênica é que a definição destes agrupamentos de genes é independente do estado específico das células em um dado momento, ou do protocolo usado para prepará-las. O ordenamento dos genes em uma lista proposto nesta invenção define uma métrica matemática que correlaciona a distância entre dois genes na lista e a probabilidade de que a ação de um influencie a ação do outro. Neste sentido, a probabilidade que dois genes interajam diminui com a distância entre suas localizações na lista ordenada e uma média dos níveis de expressão tomada sobre genes vizinhos nesta lista amortece flutuações espúrias e produz um perfil suave que chamamos de *TRANSCRIPTOGRAMA* [3].

As buscas na literatura científica e patentária apontaram alguns documentos relevantes para a presente invenção, os quais serão descritos a seguir.

O documento WO 2008/102825 descreve um método para classificar genes baseado no padrão de expressão desses genes, no qual uma análise de

cluster é realizada a fim de calcular as taxas de degradação e síntese aparente de produtos de transcrição gênica.

O documento US 2005/244822 descreve uma técnica de "*wetlab*" para monitorar a expressão gênica no qual pode ser obtido um perfil de dados de expressão gênica altamente detalhado, no qual plasmídeos que possuem sinais variáveis que possam ser detectados por NMR são incorporados antes dos promotores e é realizada uma medição quando a célula recebe algum estímulo.

O documento JP 2000/342299 descreve um método e dispositivo para analisar os dados de expressão gênica a partir do agrupamento de pontos de dados de forma que aqueles pontos que possuem padrões similares de expressão são agrupados.

A presente invenção difere de todos os documentos apresentados acima, pois propõe um processo para a construção de 'imagens' de expressão gênica de genomas inteiros pela produção de perfis de expressão para transcriptomas novo e inventivo, a partir do ordenamento de uma lista de genes considerando a integração da informação de associação proteína-proteína, fato não descrito e nem citado em nenhum documento encontrado.

Do que se depreende da literatura pesquisada, não foram encontrados documentos antecipando ou sugerindo os ensinamentos da presente invenção, de forma que a solução aqui proposta possui novidade e atividade inventiva frente ao estado da técnica.

Sumário da Invenção

Em um aspecto, a presente invenção proporciona novo e inventivo método, sistema e aparato de análise para dados de transcrição gênica. Preferencialmente a análise dos dados de transcrição é realizada através de um ordenamento do material biológico de determinado organismo (genes ou proteínas a eles associadas) de tal maneira que genes cujas proteínas têm maior probabilidade de estarem associadas de alguma forma estão mais próximos nesta lista ordenada, permitindo a produção de um meio de

visualização dessa interação (aqui também conhecido como "transcriptograma") demonstrando o nível de expressão de cada RNA mensageiro. A análise aqui apresentada inclui também um tratamento estatístico o qual evidencia alterações no metabolismo celular com precisão não atingida por outros métodos. A organização genômica obtida da forma aqui apresentada é independente de experimentos específicos e define módulos funcionais que podem ser associados aos termos de ontologia de genes.

É, portanto, um objeto da presente invenção um método de análise de dados gênicos compreendendo as etapas de:

a) ordenar uma lista de genes baseado na integração da informação de associação proteína-proteína;

b) projetar dados de expressão gênica sobre a lista de a), determinando uma média sobre os intervalos das ordenações (janela);

c) visualizar e/ou analisar através de transcriptogramas os dados obtidos de b).

Em uma realização preferencial, o ordenamento da lista compreende o uso de um simulador de Monte Carlo.

Em uma realização preferencial, para o cálculo do tamanho da janela é utilizado a medida de "modularidade da janela".

Em uma realização preferencial, as interações proteína-proteína estão especificadas em forma de uma matriz.

Em uma realização preferencial, o método acima é aplicado à análise de microarranjos.

É, também, um objeto da presente invenção um sistema e um aparato para análise de dados gênicos compreendendo as etapas de:

a) meios para ordenar uma lista de genes baseado na integração da informação de associação proteína-proteína;

b) meios para projetar dados de expressão gênica sobre a lista de a) compreendendo meios para determinar uma média sobre os intervalos das ordenações (janela);

c) meios para visualizar e/ou analisar os dados obtidos em b) através

de transcriptogramas.

Estes e outros objetos da invenção serão imediatamente valorizados pelos versados na arte e pelas empresas com interesses no segmento, e serão descritos em detalhes suficientes para sua reprodução na descrição a seguir.

Breve Descrição das Figuras

Figura 1. Matrizes de interação relativas a *Homo sapiens, Gallus gallus Saccharomyces cerevisiae,* para o ordenamento aleatório inicial (Figs. 1a-c), e após o ordenamento descrito acima (Figs. 1d-f). Para cada figura os eixos vertical e horizontal dão a posição relativa dos genes na lista, inicial (a-c) ou final (d-f). Os fundos cinza representam o perfil de modularidade versus os genes da lista. Aqui a janela é $w + 1 = 251$.

Figura 2. Matrizes de interação relativas a *Homo sapiens, Gallus gallus Saccharomyces cerevisiae,* para o ordenamento aleatório inicial (Figs. 2a-c), e após o ordenamento descrito acima (Figs. 2d-f). Para cada figura os eixos vertical e horizontal dão a posição relativa dos genes na lista, inicial (a-c) ou final (d-f). Os fundos cinza representam o perfil de modularidade versus os genes da lista. Aqui a janela é $w + 1 = 101$.

Figura 3. Perfis da distribuição dos genes pertencentes as diferentes termos da *Gene Ontology: Biological Processes,* suavizados sobre janelas de $w = 251$, referentes a *Homo sapiens* **(a-d)** , *Gallus gallus* **(e-h) e** *Saccaromyces cervisiae* **(i-l**). O fundo cinza em cada gráfico é a modularidade de janela, com $w = 251$. Observe que para todos os três organismos os picos da distribuição dos termos da GO sucedem-se com lógica biológica: os da esquerda associados à produção de biomoléculas e à direita os termos associados ao metabolismo de energia.

Figura 4. Evolução da função custo à medida que evolui o processo de ordenamento, para Homo sapiens **(a)**, *Gallus gallus* **(b)** e *Saccharomyces cerevisiae* **(c).** Cada degrau nas figuras representa a redução do parâmetro T, usado pela técnica computacional de *simulated annealing.* Cada gráfico

apresenta diferentes rodadas da simulação, escolhe-se o resultado com menor função custo.

Figura 5. Perfis da distribuição dos genes pertencentes as diferentes termos da *Gene Ontology: Biological Processes,* suavizados sobre janelas de $w = 251$, referentes a *Saccharomyces cerevisiae.* O fundo cinza em cada gráfico é a modularidade de janela, com $w = 251$. Observe que para todos os três ordenamentos, independentemente de pequenas diferenças, os picos sucedem-se com a mesma lógica biológica: os da esquerda associados à produção de biomoléculas e à direita os termos associados ao metabolismo de energia.

Figura 6. Perfis da distribuição dos genes pertencentes as diferentes termos da *Gene Ontology: Biological Processes,* suavizados sobre janelas de $w = 251$ **(a)**, $w = 101$ **(b)** e $w = 51$ **(c)**, referentes a *Saccharomyces cerevisiae.* O fundo cinza em cada gráfico é a modularidade de janela, com o tamanho de janela de acordo com a janela de suavização dos perfis da GO. *Transcriptogramas* para o *Saccharomyces cerevisiae* suavizados sobre janelas de $w = 251$ **(d)**, $w = 101$ **(e)** e $w = 51$ **(f).**

Figura 7. *Transcriptogramas* em replicata de um estado controle **(a)** e de um estado modificado **(b)** de Saccharomyces cerevisiae, como mostrados na Ref. [3]. Estes experimentos consideram amostras de levedura em estágios diferentes do ciclo respiratório. *Transcriptograma relativo* de cada replicata do controle **(c)** e de cada replicata do estado modificado **(d).** A janela utilizada para a suavização dos perfis de transcrição é de $w = 251$, e as faixas coloridas representam desvios da média do estado controle entre 0 e 2 (faixa amarela) ou entre 2 e 4 (faixa cor-de-rosa) desvios amostral (σ_s) das replicatas do estado controle, calculadas como no Caso 1 de tratamento estatístico. Para janelas de $w = 251$, no teste t de Student monocaudal um desvio de $2\sigma_s$ corresponde a valor de P=0.0233, enquanto que um desvio de $4\sigma_s$ corresponde a P=0.00005.

Figura 8. Gráfico da concentração de oxigênio dissolvido na cultura de levedura, com os dados do artigo de Tu *et al.* [8], na forma apresentada na ref.

[3]. Os pontos coloridos representam os instantes quando foram obtidos os transcriptomas para os quais produzimos *transcriptogramas relativos.*

Figura 9. *Transcriptogramas relativos* para sucessivos estados do ciclo respiratório da levedura, obtidos com dados do trabalho de Tu *et. al.* Consideramos os três estados iniciais como a amostra controle (Tempo= 0, 300 e 600 min). (a) *transcriptograma relativo* para o primeiro estado dos três ciclos, (b) para o segundo estado dos três ciclos e assim por diante. Aqui janela de suavização é $w = 251$. Para janelas de $w = 251$, no teste t de Student monocaudal um desvio de t=4 corresponde a valor de $P < 5 \times 10^{-5}$, enquanto que um desvio de t=6 corresponde a $P < 5 \times 10^{-9}$. Observe o intervalo localizado nas posições entre 0.35 e 0.45, que variam significativamente em durante o ciclo respiratório (principalmente nos tempos T=150, 450 and 4750 min, and T=175, 475 and 775 min)

Figura 10. *Transcriptogramas relativos* para sucessivos estados do ciclo respiratório da levedura, obtidos com dados do trabalho de Tu *et. al.* Consideramos os três estados iniciais como a amostra controle (Tempo= 0, 300 e 600 min). (a) *transcriptograma relativo* para o primeiro estado dos três ciclos, (b) para o segundo estado dos três ciclos e assim por diante. Aqui janela de suavização é $w = 101$. Para janelas de $w = 101$, no teste t de Student monocaudal um desvio de t=4 corresponde a valor de $P < 5 \times 10^{-5}$, enquanto que um desvio de t=6 corresponde a $P < 5 \times 10^{-9}$.

Figure 11. Oscilações nos níveis de expressão dos genes que apresentam maior variação durante o ciclo respiratório. Tu et al. reconheceram 3 grupos de 40 genes: Ox (oxidativo), R/B (redutivo, montagem), e R/C (redutivo, carregando), juntamente com os 40 genes com maior variação de expressão no intervalo 0.35-0.45 *dos transcriptogramas relativos* (amarelo). Este gráfico apresenta a média dos níveis de expressão de cada grupo de genes ao longo do ciclo respiratório.

Figura 12. *Transcritpogramas relativos* para *Saccharomyces cerevisiae*: comparação entre o metabolismo da linhagem selvagem e do mutante sgs1. (**a**) *Transcriptogramas* para duas replicatas da linhagem

selvagem e o mutante sgs1 antes da adição de MMS. (**b**) *Transcriptogramas relativos* para duas replicatas da linhagem selvagem depois do tratamento com MMS. (**c**) *Transcriptogramas relativos* para duas replicatas do mutante sgs1 depois do tratamento com MMS. (**d**) *Transcriptogramas relativos* de uma replicata da linhagem selvagem e uma replicata do mutante sgs1 depois do tratamento com MMS, para evidenciar que ambas amostras foram presas no mesmo estágio do ciclo celular. **(f)** Mesmo que **(d)**, mas para o outro par de replicatas, que foram presas em um outro estágio do ciclo celular. Todos *transcriptogramas* são relativos ao transcriptograma médio das replicatas da linhagem selvagem, antes da adição do MMS. Janela para suavização usada foi $w = 251$.

Figura 13. Classificação de Hamburguer-Hamilton para as fases do desenvolvimento embrionário de *Gallus gallus*.

Figura 14. *Transcriptogramas relativos* para as fases de Hamburguer-Hamilton tomados em relação ao *transcriptograma* médio da fase HH1.

Figura 15. Classificação dos tecidos com base nos *transcriptogramas* para 4 tecidos do *Gallus gallus*, com amostras tiradas de pintos com 0 dias **(c)** e com 7 dias **(d)**. Classificação das amostras com base na similaridade dos *transcriptogramas* **(b):** observe que os grupos mais similares são compostos pelas replicatas de um mesmo tecido, na mesma idade. A seguir são separados por idade e então por tecido.

Figura 16. *Transcriptogramas relativos* para 4 tecidos diferentes de Gallus galus, para pintos com 0 e 7 dias, tomando como referência o *transcriptograma* médio de cada tecido no dia 0. As alterações são claramente visíveis.

Figura 17. *Transcriptogramas relativos* com $w = 251$ em pacientes com adenocarcinoma de pulmão, tomados em relação ao *transcriptograma* médio dos tecidos sadios de pacientes não fumantes. Os gráficos referem-se a tecidos saudáveis (a) e cancerosos (b) de pacientes não fumantes; a tecidos

saudáveis (c) e cancerosos (d) de pacientes ex-fumantes; a tecidos saudáveis (e) e cancerosos (f) de pacientes não fumantes.

Figura 18. *Transcriptogramas relativos* com $w = 101$ em pacientes com adenocarcinoma de pulmão, tomados em relação ao *transcriptograma* médio dos tecidos sadios de pacientes não fumantes. Os gráficos referem-se a tecidos saudáveis (a) e cancerosos (b) de pacientes não fumantes; a tecidos saudáveis (c) e cancerosos (d) de pacientes ex-fumantes; a tecidos saudáveis (e) e cancerosos (f) de pacientes não fumantes.

Descrição Detalhada da Invenção

Os exemplos aqui mostrados têm o intuito somente de exemplificar uma das inúmeras maneiras de se realizar a invenção, contudo, sem limitar o escopo da mesma.

Ordenamento da lista de genes

O "ordenamento da lista de genes" da presente invenção compreende, em uma realização preferencial, o método de ordenamento descrito a seguir.

O ponto de partida para o ordenamento é uma lista de genes aleatoriamente enumerada e uma matriz de interações ou associações proteína-proteína. Consideram-se interações proteína-proteína como a associação física ou funcional apresentada por algum par de produtos gênicos.

Este corpo de informação foi produzido durantes anos por diferentes pesquisadores e laboratórios e está organizado e publicamente disponível no banco de dados STRING [2]. Tomam-se todas as associações proteína-proteína descritas neste banco de dados como "evidências experimentais" ou vindas de outros bancos de dados para um dado organismo.

Para uma lista ordenada com N genes, os dados sobre interação podem ser organizados em uma matriz M, de dimensões $N \times N$. Os elementos desta matriz, M_{ij}, com $0 \leq i \leq N$ e $0 \leq j \leq N$, assumem o valor 1 ou 0 dependendo da existência ou não de interação entre os genes localizados na i-ésima e j-ésima posições do ordenamento. O resultado é uma matriz simétrica de zeros e uns com diagonal nula. Nesta invenção propõe-se um algoritmo que

favorece a proximidade de genes interagentes minimizando uma função custo E calculada para cada ordenamento e definida como [3]

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

onde |..| garante o valor positivo para a diferença dos elementos de matriz localizados em sítios vizinhos na matriz e d_{ij} é proporcional à distância do ponto (i,j) até a diagonal, isto é, $d_{ij} = |i-j|$. Esta função custo cresce com o número de interfaces entre elementos um e zero da matriz e cresce ainda mais quando estas interfaces estão longe da diagonal. Pontos (i,j) longe da diagonal apresentam valores i e j muito diferentes, correspondendo portanto a interações entre genes distantes na lista ordenada.

Tendo como início a lista de genes aleatoriamente ordenada e a matriz de interações correspondente, o algoritmo prossegue escolhendo aleatoriamente um par de genes e tentativamente trocando suas posições no ordenamento. Uma nova matriz de interações é produzida para este ordenamento modificado e a função custo é recalculada usando a Eq.(1). Se a função custo diminui com a troca de posição de genes, essa troca é aceita. Se a função custo aumenta por ΔE, a troca é aceita com probabilidade $\exp[-\Delta E/T]$, onde T é uma temperatura virtual. No caso de $\Delta E = 0$, se aceita a troca com 50% de probabilidade. O processo é então repetido pela escolha de um novo par de genes. Inicia-se o processo com $T = 6 \times 10^5$ e a cada 100 Monte Carlo Steps (MCS) este parâmetro é reduzido de 20% de seu valor corrente. Um MCS equivale a um número de escolhas aleatórias igual ao número de elementos no sistema. Este procedimento é conhecido como *simulated annealing* [4], e visa evitar estados metaestáveis. Quando uma mudança não é aceita, todas as modificações são descartadas e um novo par de genes é escolhido. Este processo é finalizado quando o valor da função custo se estabiliza.

A Figura 1 mostra as matrizes de interação relativas a *Homo sapiens, Gallus gallus Saccharomyces cerevisiae,* para o ordenamento aleatório inicial (Figs. 1a-c), e após o ordenamento descrito acima (Figs. 1d-f). Para cada figura os eixos vertical e horizontal dão a posição relativa dos genes na lista, inicial (a-c) ou final (d-f). Estas posições estão normalizadas, de tal maneira que ao i-ésimo gene de cada lista é atribuída a posição $^{i}/_{N}$ tanto no eixo vertical como horizontal, com N sendo o número de genes na lista de cada organismo. Nestas figuras um ponto escuro localizado na posição $\left(^{i}/_{N}, ^{j}/_{N}\right)$ indica que há associação entre os genes das posições i e j da lista de genes, de tal maneira que $M_{i,j} = 1$. Cada par de configurações correspondentes ao mesmo organismo (Figs. 1a e 1d, 1b e 1e, 1c e 1f) representa a mesma informação sobre genes e associação protéica. As diferenças nas matrizes de associação protéica vêm da localização distinta dos genes sobre os eixos antes e depois de aplicado o algoritmo de ordenamento acima descrito. O resultado mostra que, enquanto antes do ordenamento os pontos escuros estão distribuídos igualmente sobre toda a figura, após o ordenamento estes estão concentrados próximos à diagonal, deixando os cantos superior esquerdo e inferior direito livres. Estes dois cantos representam associações proteína-proteína para genes localizados em pontos distantes da lista, já que eles correspondem a elementos da matriz M_{ij} para os quais i e j são muito diferentes. Mais ainda, há a formação de agrupamentos de pontos pretos perto da diagonal, que representam módulos de genes mutuamente interagentes.

Na Figura 1 a informação sobre a associação proteína-proteína foi tirada do banco de dados STRING (8ª versão) usando as evidências rotuladas como "experimental" and "database" (95% de todas associações) adicionadas com "neighbourhood", "fusion", "co-expression" e "co-occurrence" evidences, O nível do score foi ≥0.800, o que resultou em 9019 genes e 111602 associações para o *Homo sapiens*, 3850 genes e 63615 associações para o

Gallus gallus e 4655 genes e 47415 associações para o *Saccharomyces cerevisiae*.

Modularidade da janela

A "modularidade da janela" da presente invenção compreende, em uma realização preferencial, a medida da modularidade da janela descrita a seguir.

Estes módulos de pontos escuros na representação da matriz de associação protéica têm significado biológico, no sentido de que os genes cujas interações estão representadas nestes módulos *i)* interagem com maior probabilidade com genes que estão próximos deles no ordenamento e *ii)* participam ou da mesma função biológica, ou em funções relacionadas.

Com intuito de quantificar esta modularidade, propomos uma medida, a *modularidade de janela* [3], que definimos da seguinte maneira. Para cada gene da lista ordenada, considere seus $w/2$ vizinhos à esquerda e $w/2$ vizinhos à direita, compreendendo um intervalo de $w+1$ genes. A modularidade de janela, $W_w(i)$, associada ao gene na $i-$ésima posição da lista, é definida como a razão entre o número de interações que associam quaisquer dois genes na janela de tamanho $w+1$, centrada no $i-$ésimo gene, e o número de interações envolvendo pelo menos 1 gene da janela [3]. Isto é,

$$W_w(i) = \frac{2}{\sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)} \sum_{k=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)} M_{k,j}} \sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)} M_{i,j} \quad , \quad (4)$$

onde

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

dá conta das condições periódicas de contorno no caso de genes perto das pontas da lista.

Modularidade de janela depende fortemente do tamanho w da janela. Por exemplo, para uma janela do tamanho da lista ordenada, a modularidade é um para todos os genes. Por outro lado, quando o gene está centrado em um

intervalo que descreve um aglomerado de genes com muitas intra-interações, a modularidade de janela deste gene é menor se a janela for menor que o tamanho da janela, devido a interações que conectam genes dentro com genes fora da janela, mas ainda pertencentes ao aglomerado. Também, genes que se ligam a diferentes aglomerados apresentam baixa modularidade de janela. Nas Figuras 1 os perfis de modularidade de janela estão representados como fundos cinza. Nestas figuras escolhemos $w + 1 = 251$. A Figura 2 apresenta os mesmos gráficos para $w + 1 = 101$. A escolha do tamanho da janela w depende da acurácia desejada para os picos de modularidade. Na verdade, à medida que o tamanho da janela cresce, a rugosidade dos perfis de modularidade varia. Primeiramente essa rugosidade aumenta, atinge um máximo e volta a decrescer. A rugosidade de um perfil pode ser rigorosamente definida como o desvio padrão da altura dos perfis e dá uma medida da quantidade de picos e vales [5]. Na Figura 1 escolhemos $w + 1 = 251$ para dar uma descrição mais global dos termos da *Gene Ontology: Biological Processes* (mostrada a seguir) [6]. No entanto, janelas menores podem aumentar a precisão dos perfis de modularidade e dos perfis de expressão, que serão introduzidos no que segue. Observe nas figuras 1 e 2 que os perfis de modularidade mudam drasticamente antes e depois do ordenamento. Isso ocorre porque os genes próximos na lista ordenada têm maior probabilidade de pertencer a um conjunto de genes mutuamente interagentes. Tais módulos têm significado biológico.

Para mostrar este ponto, lançamos mão de uma classificação de genes feita com base nos processos biológicos dos quais eles participam. Essa informação está organizada em bancos de dados públicos, em especial o *Gene Ontology (GO) Database* [6]. Tomando a lista de genes associada a cada agrupamentos de pontos pretos, é possível obter os termos da classificação *Gene Ontology: Biological Process* que melhor descrevem aquele conjunto de genes. A seguir, é possível calcular um perfil para cada termo que é então projetado sobre a lista de genes. Isso significa tomar a lista de gene e associar a cada gene o valor 1 ou 0 dependendo do gene participar ou não do termo da

GO. Pode-se então criar um gráfico com essa informação colocando-se no eixo das abscissas a posição relativa de cada gene na lista ordenada e, no eixo das ordenadas, os valores indicativos da participação ou não de cada gene no termo. Estes perfis brutos podem ser suavizados tomando a média do perfil bruto sobre intervalos adequados. Esta média consiste em, para cada gene, fazer a média do perfil bruto sobre $w/2$ genes à sua direita e $w/2$ genes à sua esquerda, o que totaliza um conjunto de $w+1$ genes (contando com o gene central). O valor de w varia com o grau de detalhe desejado: quanto menor w, mais detalhes.

A Figura 3 apresenta diversos termos da *GO: Biological Process* projetados sobre as listas de genes de *Homo sapiens* (3a-3d), *Gallus gallus* (3e-3h) e *Saccharomyces cerevisiae* (3i-3l), antes e depois de ordená-las segundo o método proposto nesta invenção. Nestas figuras pode-se observar que *i)* antes do ordenamento os diversos termos da GO estão espalhados sobre as listas de genes, enquanto que *ii)* após ordenadas os perfis dos diversos termos das GO apresentam picos localizados. Um pico neste tipo de gráfico significa que na região do pico concentram-se os genes associados ao processo biológico representado pelo termo da GO. A mudança de perfis achatados para picos implica que os genes associados ao mesmo processo biológico foram concentrados em um mesmo intervalo da lista.

O ordenamento e a projeção dos termos da GO nas listas ordenadas possibilita a identificação de intervalos da lista ordenada com funções importantes do metabolismo celular. Na verdade, a ordem dos picos dos termos da GO reproduz o ciclo celular: a primeira metade destas listas apresenta picos de termos relacionados metabolismo de biomoléculas, enquanto que a segunda metade está associada com metabolismo energético. Mais detalhes podem ser observados diretamente na Fig. 3.

Transcriptogramas

Os transcriptogramas da presente invenção são preferencialmente descritos a seguir.

A produção dos *transcriptogramas*

A produção dos *transcriptogramas* requer informação sobre expressão gênica. Usualmente isso significa dados de micro-arranjos. No banco de dados *Gene Expression Omnibus* [7] estão disponíveis dados de experimentos com micro-arranjos com sondas para quase todos componentes conhecidos dos genomas de diferentes organismos. Um *transcriptograma* é obtido associando-se a cada gene do ordenamento o valor da sua expressão e, em seguida, calculando a média dessas expressões sobre janelas de tamanho adequadamente escolhido. Em outras palavras, os *transcriptogramas* são obtidos associando ao i-ésimo gene do ordenamento a média dos valores de expressão dos genes em uma janela de tamanho $w + 1$, centrada naquele gene.

A base desta invenção é o ordenamento de genes que produz uma lista onde genes próximos têm maior probabilidade de apresentar produtos gênicos interagentes ou associados ao mesmo processo biológico. Neste caso a proximidade na lista dá um critério de vizinhança para o processo de suavização de diversos perfis.

O controle do processo de ordenamento é feito com base no valor da função custo E dada pela Eq.(1). À medida que transcorre o processo de ordenamento via Monte Carlo, monitora-se a função custo. A Fig.4 apresenta a evolução de E durante este processo para diferentes organismos e diferentes sequências de números aleatórios. Nota-se que a função decresce monotonicamente, embora a cada mudança do parâmetro T, a temperatura virtual do '*simulated annealing*', E desce um degrau. Os ordenamentos estão prontos quando E estabiliza. Diferentes rodadas do programa para um mesmo organismo podem levar a diferentes ordenamentos. Essas diferenças, no entanto, não são importantes, no sentido que não diferem muito na aproximação dos genes associados ao mesmo processo biológico, nem na sequência com que estes processos estão apresentados. Veja a Fig.5, onde apresentamos as projeções para termos da GO em três diferentes

ordenamentos do *Saccharomyces cerevisiae* obtidos em três rodadas do método (com sequências de números aleatórios diferentes).

Um segundo parâmetro de controle é o tamanho da janela usado tanto para as médias do enriquecimento dos termos da *Gene Ontology,* como dos *transcriptogramas.* O tamanho mais adequado para as janelas depende do grau de detalhe desejado: janelas grandes dão uma visão mais global, enquanto que janelas menores ressaltam mais detalhes. A Fig.6 apresenta as projeções de termos associados a processos biológicos e um *transcriptograma* para *Saccharomyces cerevisiae* com diferentes janelas.

Etapas para a produção dos Transcriptogramas

Uma maneira de analisar alterações nos níveis de expressão de diferentes experimentos com relação a um controle considera as seguintes etapas:

1) Produz-se o *transcriptograma* para todas as replicatas do experimento-controle, com uma janela escolhida adequadamente. Isto significa que se toma os valores de expressão τ_i^α associados ao gene localizado na i-ésima posição da α-ésima das n replicatas dos transcriptomas que compõem a amostra controle, e obtém-se os valores 'janelados' , $\langle \tau_i^\alpha \rangle_w$, que são médias sobre intervalos de tamanho $w+1$, isto é

$$\langle \tau_i^\alpha \rangle_w = \frac{1}{w+1} \sum_{j=\mathrm{mod}\left(i-\frac{w}{2},N\right)}^{\mathrm{mod}\left(i+\frac{w}{2},N\right)} \tau_j^\alpha \qquad (5)$$

2) Para cada gene, calcula-se a média dos valores do *transcriptograma* das replicatas do controle:

$$\langle \tau_i^0 \rangle_w = \frac{1}{n} \sum_{\alpha=1}^{n} \langle \tau_i^\alpha \rangle_w \qquad (6)$$

3) Tomam-se então os *transcriptogramas* das replicatas individualmente, do controle e das amostras que não são controle e divide-se o valor de expressão pelo valor da média dos controles. Esta etapa produz um

transcriptograma relativo ao controle, onde valores acima ou abaixo de 1 significam, respectivamente, aumento ou diminuição dos níveis de expressão em relação à média.

Tratamento estatístico dos *transcriptogramas*

A vantagem do *transcriptograma* é a redução da flutuação aleatória por meio da média dos valores de expressão tomados sobre intervalos de tamanho $w+1$, valor então atribuído ao gene central da janela. Correndo a janela um gene para a direita tem-se então um novo conjunto de genes e uma nova média pode ser calculada.

O tratamento estatístico para o *transcriptograma* é feito para cada janela e tratando cada uma como uma amostra de $w+1$ genes. A hipótese nula é que o valor médio de expressão de cada janela é igual à média de expressão sobre a mesma janela para o conjunto de medidas do controle. Observe que desta maneira ficam dispensadas as replicatas, embora claro, replicatas tendem a melhorar as estimativas.

Caso 1: Vamos considerar cada valor janelado como uma medida. Tendo n replicatas de um estado controle, consideramos como hipótese que o valor médio de expressão de cada janela é igual à média de expressão sobre a mesma janela para a média do conjunto de medidas do controle. Neste caso, seja $\langle \tau_i^{\alpha} \rangle_w$ o nível de expressão média dos genes localizados na janela de tamanho $w+1$ centrada no i-ésimo gene de um transcriptoma , com $\alpha = 1, 2, \cdots, n$ designando cada replicata do estado controle. Seja $\langle \tau_i^{b} \rangle_w$ o nível médio de expressão da mesma janela do outro transcriptoma. Então os valores médios (janelados) para cada janela podem ser calculados como

$$\langle \tau_i^{\alpha} \rangle_w = \frac{1}{w+1} \sum_{j=\mathrm{mod}\left(i-\frac{w}{2},N\right)}^{\mathrm{mod}\left(i+\frac{w}{2},N\right)} \tau_j^{\alpha} \quad , \qquad (5)$$

e podemos calcular a média sobre as replicatas do controle,

$$\overline{\langle \tau_i^{0} \rangle}_w = \frac{1}{n} \sum_{\alpha=1}^{n} \langle \tau_j^{\alpha} \rangle_w \quad , \qquad (6)$$

bem como a média dos valores quadráticos,

$$\overline{\langle \tau_j^{\alpha} \rangle_w^2} = \frac{1}{n} \sum_{\alpha=1}^{n} \langle \tau_j^{\alpha} \rangle_w^2 \quad , \qquad (7)$$

de tal maneira que podemos estimar o desvio padrão amostral, como

$$\sigma_i^0 = \sqrt{\frac{\overline{\langle \tau_j^{\alpha} \rangle_w^2} - \overline{\langle \tau_i^0 \rangle_w}^2}{n-1}} \quad . \qquad (8)$$

O desvio padrão amostral dá uma estimativa da variação dos valores dos transcriptogramas tomados em janelas de tamanho $w+1$. Quando o valor de outro transcriptograma, digamos da amostra b, dado por

$$\langle \tau_i^b \rangle_w = \frac{1}{w+1} \sum_{j=\mathrm{mod}\left(i-\frac{w}{2},N\right)}^{\mathrm{mod}\left(i+\frac{w}{2},N\right)} \tau_j^b \quad , \qquad (9)$$

se afasta da média das amostras, $\overline{\langle \tau_i^0 \rangle_w}$, por vários desvios padrões amostrais, σ_i^0, pode-se dizer que esta diferença é significativa. Desta maneira ficam dispensadas as replicatas, embora claro, replicatas tendem a melhorar as estimativas.

Caso 2: O tratamento estatístico para o *transcriptograma* pode ser feito para cada janela tratando-as como amostras de $w+1$ medidas. A hipótese nula é que o valor médio de expressão de cada janela é igual à média de expressão sobre a mesma janela para o conjunto de medidas do controle. Observe que desta maneira ficam dispensadas as replicatas, embora claro, replicatas tendem a melhorar as estimativas.

O caso 2 refere-se à probabilidade de que o valor janelado de um transcriptograma é significantemente diferente do valor janelado de outro transcriptograma, levando em conta o tamanho da janela.

Seja τ_i^a o nível de expressão do i-ésimo gene de um transcriptoma e τ_i^b o nível de expressão do i-ésimo gene do outro transcriptoma. Então os valores médios (janelados) e o desvio padrão para cada janela podem ser calculados como

$$\langle \tau_i^{a,b} \rangle = \frac{1}{w+1} \sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)} \tau_j^{a,b} \quad , \qquad (10a)$$

$$\langle (\tau_i^{a,b})^2 \rangle = \frac{1}{w+1} \sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)} (\tau_j^{a,b})^2 \quad , \qquad (10b)$$

$$\sigma_i^{a,b} = \sqrt{\langle (\tau_i^{a,b})^2 \rangle - \langle \tau_i^{a,b} \rangle^2} \quad , \qquad (10c)$$

onde

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

para lidar com os genes nas pontas da lista ordenada (condições de contorno periódicas). $\sigma_i^{a,b}$ é desvio padrão para a janela de tamanho $w+1$ centrado no gene da i-ésima posição do ordenamento da amostra a ou b.

Definimos agora os parâmetros t_i e df_i (número de graus de liberdade), referentes à janela centrada no i-ésimo gene, usados em teste-t usuais para a determinação dos intervalos de significância:

$$t_i = \frac{\langle \tau_i^a \rangle - \langle \tau_i^b \rangle}{s_i^{ab}}, \qquad (11a)$$

onde

$$s_i^{ab} = \sqrt{\frac{(\sigma_i^a)^2 + (\sigma_i^b)^2}{w+1}}, \qquad (11b)$$

e

$$df_i = w \frac{[(\sigma_i^a)^2 + (\sigma_i^b)^2]^2}{(\sigma_i^a)^4 + (\sigma_i^b)^4}. \qquad (11c)$$

Caso 3: Probabilidade de que o valor janelado de um transcriptograma é significantemente diferente do valor janelado da média de n replicatas de transcriptogramas.

Seja τ_i^a o nível de expressão do i-ésimo gene do a-ésimo transcriptoma de um conjunto de n replicatas e τ_i^b o nível de expressão do i-ésimo gene do outro transcriptoma. Então, os valores associados ao conjunto das n replicatas podem ser calculados como

$$\langle \tau_i^b \rangle = \frac{1}{n(w+1)} \sum_{a=1}^{n} \sum_{j=\mathrm{mod}\left(i-\frac{w}{2},N\right)}^{\mathrm{mod}\left(i+\frac{w}{2},N\right)} \tau_j^a \, , \qquad (12a)$$

$$\langle(\tau_i^0)^2\rangle = \frac{1}{n(w+1)}\sum_{\alpha=1}^{n}\sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)}\left(\tau_j^\alpha\right)^2 . \qquad (12b)$$

$$\sigma_i^0 = \sqrt{\langle(\tau_i^0)^2\rangle - \langle\tau_i^0\rangle^2} . \qquad (12c)$$

Definimos agora os parâmetros τ_i e df_i, usados em teste-t para a determinação dos intervalos de significância:

$$t_i = \frac{\langle\tau_i^0\rangle - \langle\tau_i^b\rangle}{s_i^{0b}} , \qquad (13a)$$

onde

$$s_i^{0b} = \sqrt{\frac{(\sigma_i^0)^2}{n(w+1)} + \frac{(\sigma_i^b)^2}{w+1}} , \qquad (13b)$$

e

$$df_i = w[n(w+1)-1]\frac{\left[\frac{(\sigma_i^0)^2}{n(w+1)} + \frac{(\sigma_i^b)^2}{w+1}\right]^2}{w\left[\frac{(\sigma_i^0)^2}{n(w+1)}\right]^2 + n(w+1)\left[\frac{(\sigma_i^b)^2}{w+1}\right]^2} . \qquad (13c)$$

Caso 4: Probabilidade que o valor janelado da média de n replicatas de transcriptogramas seja significantemente diferente da média de outras n replicatas de transcriptogramas.

Seja τ_i^α o nível de expressão do i-ésimo gene do α-ésimo transcriptoma de um conjunto de n replicatas e τ_i^β o nível de expressão do i-ésimo gene do β-ésimo transcriptoma de outro conjunto de n replicatas. Então, os valores associados a cada conjunto das n replicatas podem ser calculados como

$$\langle\tau_i^0\rangle = \frac{1}{n(w+1)}\sum_{\alpha=1}^{n}\sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)}\tau_j^\alpha , \qquad (14a)$$

$$\langle(\tau_i^0)^2\rangle = \frac{1}{n(w+1)}\sum_{\alpha=1}^{n}\sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)}\left(\tau_j^\alpha\right)^2 , \qquad (14b)$$

$$\sigma_i^0 = \sqrt{\langle(\tau_i^0)^2\rangle - \langle\tau_i^0\rangle^2} , \qquad (14c)$$

$$\langle \tau_i^1 \rangle = \frac{1}{n(w+1)} \sum_{\beta=1}^{n} \sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)} \tau_j^\beta \,, \qquad (14d)$$

$$\langle (\tau_i^1)^2 \rangle = \frac{1}{n(w+1)} \sum_{\beta=1}^{n} \sum_{j=\mathrm{mod}(i-w/2,N)}^{\mathrm{mod}(i+w/2,N)} \left(\tau_j^\beta\right)^2 \,, \qquad (14e)$$

$$\sigma_i^1 = \sqrt{\frac{\langle (\tau_i^1)^2 \rangle - \langle \tau_i^1 \rangle^2}{n(w+1)-1}} \,. \qquad (14f)$$

Definimos agora os parâmetros t_i e df_i, usados em teste-t para a determinação dos intervalos de significância:

$$t_i = \frac{\langle \tau_i^0 \rangle - \langle \tau_i^1 \rangle}{s_i^{01}} \,, \qquad (15a)$$

onde

$$s_i^{01} = \sqrt{\frac{(\sigma_i^0)^2 + (\sigma_i^1)^2}{n(w+1)}} \,, \qquad (15b)$$

e

$$df_i = (n(w+1)-1)\frac{[(\sigma_i^0)^2 + (\sigma_i^1)^2]^2}{(\sigma_i^0)^4 + (\sigma_i^1)^4} \,. \qquad (15c)$$

Com os valores que t_i e df_i calculados adequadamente para cada caso, torna-se possível a aplicação do teste-t usual. Observe que t_i e df_i dependem do subscrito i, significando que o teste de significância depende da posição da janela, gerando assim bandas de significância para os *transcriptogramas.*

No caso específico do ***transcriptograma relativo*** de janela $w+1$, os níveis de transcrição para a janela centrada no gene localizado na posição i são divididos por $\langle \tau_i^0 \rangle$. Neste caso, tanto os valores médios quanto os desvios padrão são dados em unidades de $\langle \tau_i^0 \rangle$, de maneira que os valores de t_i e df_i ficam iguais.

Com os desvios padrão amostral do valor associado a cada gene pelo *transcriptograma,* tornam-se possíveis gráficos como os apresentados na Fig. 7. Apresentamos ali uma estimativa da variação dos valores pelo Caso 1.

Escolhemos pintar duas faixas: amarela, que representa pontos que se desviam da média dos controles entre 0 a 2 desvios padrões das médias dos controles, e cor-de-rosa, que representa pontos que desviam das médias entre 2 e 4 desvios padrões das médias dos controles. Esses valores podem ser traduzidos em valores P de confiança, que, claro, dependem do tamanho da amostra. (Em testes t de Student monocaudal , para $w = 250$, $P(t = 2) < 0.0233$ e $P(t = 4) < 0.00005$. Para $w = 100$, $P(t = 2) < 0.0241$ e $P(t = 4) < 0.00005$.)

Sobre o gráfico produzido com base nas médias dos controles e seus desvios padrão amostral podem ser graficados os *transcriptogramas relativos* (isto é, divididos pelos *transcriptograma* médio dos controles). As faixas coloridas auxiliam na detecção de regiões do ordenamento associadas a genes cujas expressões estão significantemente alteradas em relação ao controle. Veja a Figura 7 c e d.

Observe que se o número de genes N cresce, cresce também o número de janelas. E fazer muitas medidas aumenta a probabilidade de obtermos algum resultado pouco provável. Assim, o intervalo de confiança deve ser tomado de tal maneira que $NP \ll 1$.

Exemplo 1. Realização Preferencial

Esta invenção consiste em um processo para ordenar uma lista de genes usando o método de física computacional conhecido como Monte Carlo [1]. O objetivo é produzir uma lista de genes onde aqueles mutuamente interagentes estejam próximos, de tal maneira que a distância entre dois genes da lista esteja correlacionada com a probabilidade que seus produtos gênicos estejam associados. O critério para associação ou não de duas proteínas pode ser obtido de bancos de dados públicos como o STRING [2]. Uma primeira vantagem em relação ao agrupamento por co-expressão gênica é que a definição destes agrupamentos de genes é independente do estado específico das células em um dado momento, ou do protocolo usado para prepará-las. O ordenamento dos genes em uma lista proposto nesta invenção define uma métrica matemática que correlaciona a distância entre dois genes na lista e a

probabilidade de que a ação de um influencie a ação do outro. Neste sentido, a probabilidade que dois genes interajam diminui com a distância entre suas localizações na lista ordenada e uma média dos níveis de expressão tomada sobre genes vizinhos nesta lista amortece flutuações espúrias e produz um perfil suave que chamamos de *TRANSCRIPTOGRAMA* [3].

Exemplo 1: Ciclo celular de levedura.

Para a o ordenamento da levedura *Saccharomyces cerevisiae,* consideramos interações entre genes ou proteínas as associações físicas ou funcionais de produtos protéicos como disponibilizado no banco de dados STRING database [2]. Tomamos todas as interações proteína-proteína descritas no STRING como inferidas por evidências ali qualificadas como "*experimental*" e "database" para o organismo *Saccharomyces cerevisiae*. Nossa lista final constitui-se de 4655 genes e 47415 interações.

Os dados de expressão que discutimos neste exemplo correspondem ao experimento relatado por Tu *et al*. [8], e foram obtidos de uma cultura controlada de levedura, onde os níveis de concentração de O_2 dissolvidos foram constantemente monitorados. Durante o experimento estes níveis variaram periodicamente e os níveis de expressão gênica foram medidos para 12 diferentes instantes durante três períodos de oscilação dos níveis de O_2 dissolvidos, formando um conjunto de 36 perfis de transcrição. A Fig.8 mostra o gráfico log-linear dos níveis de O_2 dissolvidos pelo tempo, onde estão assinalados os pontos onde foram medidos os níveis de expressão.

Nas Figuras 9 e 10 apresentamos os resultados dos *transcriptogramas relativos*, onde o estado controle foi definido como o estado inicial de cada ciclo. Assim, calculamos os *transcriptogramas* para uma dada janela, calculamos a média e o desvio padrão amostral, como descrito no Caso 1, dos três *transcriptogramas* associados com os tempos indicados como t=0, 300, e 600 min. Todos os *transcriptogramas* são então divididos pelo *transcriptograma* médio inicial. Os resultados, para janelas de 101 e 251 estão mostrados nas figuras 9 e 10. Cada gráfico nessas figures representa um instante no ciclo respiratório da cultura de levedura e, como foram medidos três ciclos, mostram

três *transcriptogramas*. A faixa amarela e cor-de-rosa refere-se a alterações em relação ao estado inicial de, respectivamente, 0 a 2 e 2 a 4 desvios padrões amostrais.

Os perfis de expressão mostram comportamentos diferentes ao longo do ordenamento: à esquerda o perfil apresenta muito abruptamente um pico associado a consumo intenso de oxigênio, enquanto que o lado direito sobe gradualmente quando as células diminuem o consumo de oxigênio. De acordo com os processos biológicos mapeados nas Fig. 6, o lado direito está associado com vários processos que demandam energia, essencialmente representados pela síntese de polímeros biológicos. Tal síntese requer grandes quantidades de adenosina trifosfatada (ATP), que está disponibilizada em profusão na fase respiratória. A alternância das rotas metabólicas para produção de energia é compatível com o ordenamento temporal através das fases em como descrito no artigo original [8].

Os resultados mostrados pelos *transcriptogramas relativos* dão suporte às conclusões tiradas por Tu *et al.* [8], baseadas na expressão dos 40 genes para cada conjunto, uma pequena fração dos genes presentes nos transcriptomas de levedura. O *transcriptograma*, por outro lado, apresenta as alterações dinâmicas durante o ciclo metabólico lançando mão da informação completa.

Mais ainda, os *transcriptogramas* possibilitam conclusões adicionais. Há mais regiões no ordenamento que variam significantemente durante o ciclo respiratório da levedura. As Figuras 9 e 10 mostram que no intervalo de 0.35 a 0.45 sobre o ordenamento há alterações significantes durante o ciclo respiratório. Veja em especial as Figuras 9 e 10 associadas aos tempos entre 25 e 125 min. Como ilustração, na Fig.11 apresentamos a média dos níveis de expressão dos três grupos de 40 genes descritos por Tu *et. al.*, juntamente com a média das expressões dos 40 genes mais alterados no intervalo entre 0.35 e 0.45 do ordenamento. Embora este último grupo oscile menos intensamente, ainda estas oscilações são altamente significativas.

O grupo de genes identificado pelo *transcriptograma relativo* é rico em genes pertencentes a processos de catabolismo de macromoléculas ou transporte nuclear. Na verdade, estes 40 genes pertencem a dois diferentes subpicos do pico 4 da Fig. 3, onde foi usada uma janela Esse maior detalhe é mostrado quando usa-se uma janela menor ao invés de, como pode ser visto na Fig. 10.

Exemplo 2: Alterações de expressão devido a tratamentos diferentes em leveduras.

Consideramos agora um experimento relatado por Fry, Sambandan, and Rha [9], onde os autores comparam níveis de transcrição de uma linhagem selvagem do *Saccharomyces cerevisiae* com linhagens mutantes com o gene sgs1 deletado. Culturas de ambas linhagens são submetidas a estresse, representado pela adição direta de 0.1% metil metanoesulfonato (MMS) e incubação a 30 °C por 1 h. As conclusões do artigo, tiradas a partir dos resultados foram *i)* em condições normais a linhagem mutante apresenta 4% dos genes com níveis transcricionais alterados em duas vezes ou mais e *ii)* sob condições de estresse não há diferenças significativas entre os níveis de expressão das linhagens selvagem e mutada. A Figura 12 apresenta os *transcriptogramas* destas amostras, relativos à média das duas replicatas das culturas selvagens sem adição de MMS. Esta figura apresenta como fundo cinza a modularidade, para guiar os olhos. Estes dados foram tirados do banco de dados *Gene Expression Omnibus*, GSE423, associados ao experimento de Fry *et al.*

A Figura 12a apresenta os *transcriptogramas relativos* para as condições sem estresse para as duas linhagens, em replicatas. Observamos primeiramente que embora não haja nenhum pico muito alto, os mutantes sgs1 dão lugar a perfis de expressão relativos com valores geralmente menores que 1, isto é, os níveis de expressão do mutantes sgs1 estão consistentemente menores que aqueles da linhagem selvagem, possivelmente indicando uma redução generalizada do metabolismo celular devido ao *knock-out* do gene sgs1. As Figuras 12b e 12c mostram *transcriptogramas relativos* para,

respectivamente, as linhagens selvagem e mutante depois do tratamento com MMS. Os *transcriptogramas relativos* tomaram novamente como referência o transcriptograma médio das replicatas da linhagem selvagem sob condições normais. Observe que cada uma destas figuras apresentam dois transcriptogramas muito diferentes. Estas diferenças são observáveis devido aos picos e depressões que representam alterações em relação ao estado normal da linhagem selvagem. No entanto, levando em consideração que os *transcriptogramas* para o ciclo respiratório apresentados nas Figuras 9 e 10, podemos supor que em cada caso as replicatas foram presas em diferentes estágios do ciclo respiratório. Na verdade, a adição de MMS pode prender as células em diferentes estágios do ciclo celular. Estes *transcriptogramas* indicam que as células de cada cultura foram presas em diferentes estágios do ciclo celular. Para evidenciar ainda mais, as Figuras 12d e 12e apresentam, cada uma, a superposição de uma replicata da linhagem selvagem e uma replicata do mutante sgs1: os gráficos são agora quase idênticos. Estes gráficos corroboram as conclusões de Fry e colaboradores que, sob estresse induzido por MMS, o metabolismo da linhagem selvagem e do mutante sgs1 apresentam performances metabólicas equivalentes. No entanto, esses gráficos também apontam que cuidado deve ser tomado naquilo que diz respeito ao estágio do ciclo celular no qual encontram-se as células da cultura, por meio de uma sincronização das células da cultura ou determinando em qual estágio estão efetivamente as células no momento da extração do mRNA.

Exemplo 3: Diferenças de expressão durante o desenvolvimento embrionário em *Gallus gallus.*

Animais superiores, como aves e mamíferos, apresentam ontogenia, onde o organismo desenvolve-se a partir da célula ovo até um animal completo. Neste desenvolvimento as células sofrem sucessivas duplicações, diferenciação e apoptose. Em especial para *Gallus gallus*, um organismo modelo para aves, as fases de desenvolvimento embrionário foram bastante descritas tendo gerada uma classificação de consenso, a classificação de Hamburger-Hamilton [10]. Como ilustração veja a figura 13.

Usando dados disponibilizados Irie e Kuratani [11], produzimos *transcriptogramas relativos* para cada duplicata referente a cada estágio de desenvolvimento embrionário, tomando como referência a média dos *transcriptogramas* das replicatas associadas ao primeiro estágio de desenvolvimento, HH1. Assim, tais perfis mostram alterações dos níveis de expressão do embrião inteiro, ao longo do desenvolvimento, em relação ao estágio HH1, como mostrados na Fig. 14. Observe que para cada estágio as replicatas dão lugar a *transcriptogramas* praticamente idênticos, enquanto que, com ajuda das projeções dos termos da GO das Fig. 3e-3h, pode-se apontar as regiões do ordenamento onde aparecem alterações de níveis de expressão à medida que se passa para estágios mais avançados de desenvolvimento. De HH1 para HH2 há um aumento na expressão na maioria dos trechos do ordenamento. Há vários picos, começando pelo intervalo do ordenamento entre as posições relativas 0.10 e 0.15 que está associado à regulação do citoesqueleto de actina, seguido por uma depressão próxima de 0.15, onde estão localizados genes ligados a adesão célula-substrato e processo baseado em filamentos de actina, na região ao redor de 0.20 há um pico de expressão ligado à migração celular e transporte de vesícula de Golgi. Um novo pico se localiza entre 0.20 e 0.25 relacionado com processos metabólicos de nucleosídeos, seguido por um pequeno pico ao redor de 0.30, relacionado com transcrição de DNA, depois uma pequena depressão pouco antes da posição 0.4, relacionado com replicação de DNA, então um pico ao redor de 0.45 relacionado com transporte de RNA e exportação nuclear. Mudando agora a natureza dos processos biológicos, ao redor de 0.55 há um pico associado com metabolismo de carbohidratos, seguidos de vários picos relativos à tradução de RNA. A região ao redor da posição 0.8, relacionada a fatores de crescimento e sinalização transmembrana está moderadamente alterada.

À medida que se passa aos *transcriptogramas relativos* de outros estágios de desenvolvimento esses padrões vão se alterando, com picos crescendo e diminuindo, fornecendo uma imagem global das alterações metabólicas sofridas pelas células do organismo durante o desenvolvimento

embrionário, mas com detalhe suficiente para que as alterações das diversas rotas metabólicas estejam ressaltadas.

Exemplo 4: Diferenças de expressão em diferentes tecidos celulares de *Gallus gallus.*

Partindo agora de outro experimento, realizado por Delfino [12], tecidos específicos foram tirados de pintos recém-saídos do ovo e após 7 dias, sempre em triplicatas. Os tecidos escolhidos foram músculo peitoral, cérebro inteiro, fígado e duodeno. Nestes casos, dado que se tomam amostras de células tiradas de um mesmo tecido, os *transcriptogramas* são notavelmente robustos para fins diagnósticos. Os dados estão disponíveis em [12]. A figuras 15c e 15d mostram os *transcriptogramas* absolutos onde o eixo das ordenadas foi normalizado para que a área debaixo dos gráficos seja um. Observe que para cada caso (tecido e idade do pinto) as triplicatas são quase indistinguíveis, mas de tecido para tecido há alterações evidentes. Para mostrar esse ponto, definimos o grau de similaridade de dois *transcriptogramas* como a área entre os dois perfis: se os *transcriptogramas* forem idênticos, tal área é zero. Medimos essa área entre *transcritpogramas*, todos versus todos e daí obtivemos uma árvore hierárquica das amostras por similaridade, na forma de um dendograma . A figura 14b mostra o resultado de tal árvore. Observe que os primeiros agrupamentos ocorrem para as triplicatas e depois por tecido. Este exemplo mostra a capacidade do *transcriptograma* de identificar tecidos semelhantes e separar os diferentes, propriedade essencial para fins diagnósticos.

Para melhor evidenciar esta capacidade diagnóstica, na Fig. 16 apresentamos *transcriptogramas relativos* onde o controle, para cada tecido, é a média dos *transcriptogramas* daquele tecido tomados dos pintos de 0 dias. A faixa amarela e cor-de-rosa referem-se a valores associados a alterações entre zero e 2 ou 2 e 4 desvios padrões da amostrais, calculados como no Caso 1, das triplicatas do dia zero de cada tecido. As alterações são claramente evidentes, sendo características de cada tecido.

Exemplo 5: Diagnóstico de câncer em tecidos pulmonares humanos.

Finalmente apresentamos um exemplo aplicado diretamente a diagnóstico. Os dados de transcrição foram novamente obtidos do banco de dados público *Gene Expression Omnibus,* da série GSE10072, que faz parte de um estudo sobre a genética e expressão de câncer [13]. Estas amostras referem-se a pacientes humanos com adenocarcinoma de pulmão. Consideramos dados de 33 pacientes, sendo que 11 são não fumantes, 10 ex-fumantes e 12 fumantes. De cada paciente foram retirados tecidos cancerosos, juntamente com tecido sadio do pulmão. Para fins diagnósticos, o *transcriptograma relativo* é o que mais evidencia as alterações em tecidos normais. Usamos como referência a média dos *transcriptogramas* de tecidos normais dos pacientes que nunca fumaram, e dessas 11 amostras calculamos também o desvio padrão amostral para cada gene. Como discutido anteriormente, esses valores dependem da janela escolhida. Os *transcriptogramas relativos* são então obtidos pela razão entre os valores do perfil janelado de expressão de cada gene pelo valor médio dos perfis janelados para as amostras de tecidos normais de pacientes não fumantes. A Figura 17 mostra esses *transcriptogramas relativos*, em 6 diferentes gráficos: tecidos sadios e cancerosos dos pacientes não fumantes, ex-fumantes e não fumantes. Observa-se claramente que os tecidos cancerosos estão alterados em relação aos tecidos normais de não fumantes, mas também podemos observar que os tecidos normais dos pacientes ex-fumantes e fumantes já apresentam alterações nos níveis de expressão. As alterações podem ser identificadas com processos biológicos, usando a caracterização biológica apresentada na Fig.3. Assim, células cancerosas de adenocarcinoma de pulmão de pacientes fumantes apresentam um aumento de atividade nas regiões do *transcriptogramas* nos intervalos entre 0.1 e 0.215, que estão associados a respiração celular e metabolismo de energia, depressão nos intervalos entre 0.31 e 0.33, 0.35 e 0.38, 0.45 e 0.5, associadas, respectivamente, ao transporte de proteínas, processos envolvendo filamentos de actina, e ao sistema imunológico. Uma nova sucessão de aumentos na intensidade de expressão agora entre 0.65 e 0.7e, região associada à mitose e

replicação e reparo de DNA e entre 0.82 e 0.9, intervalo enriquecido com genes associados à tradução de RNA. Esse *transcriptograma* é, portanto capaz de detectar em cada medida, alterações importantes que estão ligadas ao aumento de divisão celular, diminuição da atividade relacionada com o sistema imune e respostas inflamatórias, além de aumentar a produção de biomoléculas pelo aumento da atividade transcricional. Mais que isso, as faixas coloridas dão a significância das alterações. *Transcriptogramas* de janelas menores discriminam melhor os diferentes picos e depressões nos *transcriptogramas*, o que pode ser interessante na determinação, com maior especificidade, das rotas alteradas em cada exame. Veja Fig. 18, onde alguns picos e depressões estão divididos.

Os versados na arte valorizarão os conhecimentos aqui apresentados e poderão reproduzir a invenção nas modalidades apresentadas e em outros variantes, abrangidos no escopo das reivindicações anexas.

Reivindicações

MÉTODO, SISTEMA E APARATO DE ANÁLISE DE DADOS DE EXPRESSÃO GÊNICA (TRANSCRIPTOGRAMA)

1. Método de análise de dados gênicos caracterizado por compreender as etapas de:
   a) ordenar uma lista de genes baseado na integração da informação de associação proteína-proteína;
   b) projetar dados de expressão gênica sobre a lista de a), determinando uma média sobre os intervalos das ordenações (janela);
   c) visualizar e/ou analisar através de transcriptogramas os dados obtidos de b).
2. Método, de acordo com a reivindicação 1, caracterizado pelo ordenamento da lista compreender o uso de um simulador de Monte Carlo.
3. Método, de acordo com a reivindicação 1, caracterizado pelo cálculo do tamanho da janela compreender a medida de modularidade da janela.
4. Método, de acordo com a reivindicação 1, caracterizado pelo ordenamento de genes produzir uma lista onde genes próximos têm maior probabilidade de apresentar produtos gênicos interagentes ou associados ao mesmo processo biológico.
5. Método, de acordo com a reivindicação 4, caracterizado pela proximidade na lista dar um critério de vizinhança para o processo de suavização de diversos perfis.
6. Método, de acordo com a reivindicação 1, caracterizado pelas interações proteína-proteína estarem especificadas em forma de uma matriz.
7. Método, de acordo com a reivindicação 1, caracterizado pelo método ser aplicado a análise de microarranjos.

8. Sistema para análise de dados gênicos, caracterizado por compreender as etapas de:
   a) meios para ordenar uma lista de genes baseado na integração da informação de associação proteína-proteína;
   b) meios para projetar dados de expressão gênica sobre a lista de a) compreendendo meios para determinar uma média sobre os intervalos das ordenações (janela);
   c) meios para visualizar e/ou analisar os dados obtidos em b) através de transcriptogramas.
9. Aparato para análise de dados gênicos, caracterizado por compreender um sistema compreendendo as etapas de:
   a) meios para ordenar uma lista de genes baseado na integração da informação de associação proteína-proteína;
   b) meios para projetar dados de expressão gênica sobre a lista de a) compreendendo meios para determinar uma média sobre os intervalos das ordenações (janela);
   c) meios para visualizar e/ou analisar os dados obtidos em b) através de transcriptogramas.

Resumo

MÉTODO, SISTEMA E APARATO DE ANÁLISE DE DADOS DE EXPRESSÃO GÊNICA (TRANSCRIPTOGRAMA)

A presente invenção descreve um novo e inventivo método, sistema e aparato de análise para dados de transcrição gênica. Preferencialmente a análise dos dados de transcrição é realizada através de um ordenamento da informação biológica de determinado organismo (genes ou proteínas a eles associadas) de tal maneira que genes cujas proteínas têm maior probabilidade de estarem associadas de alguma forma estão mais próximos nesta lista ordenada, permitindo a produção de um meio de visualização dessa interação (aqui também conhecido como "transcriptograma") demonstrando o nível de expressão de cada RNA mensageiro. A análise aqui apresentada inclui também um tratamento estatístico o qual evidencia alterações no metabolismo celular com precisão não atingida por outros métodos.

ANEXOS

FIGURAS

Figura 1:

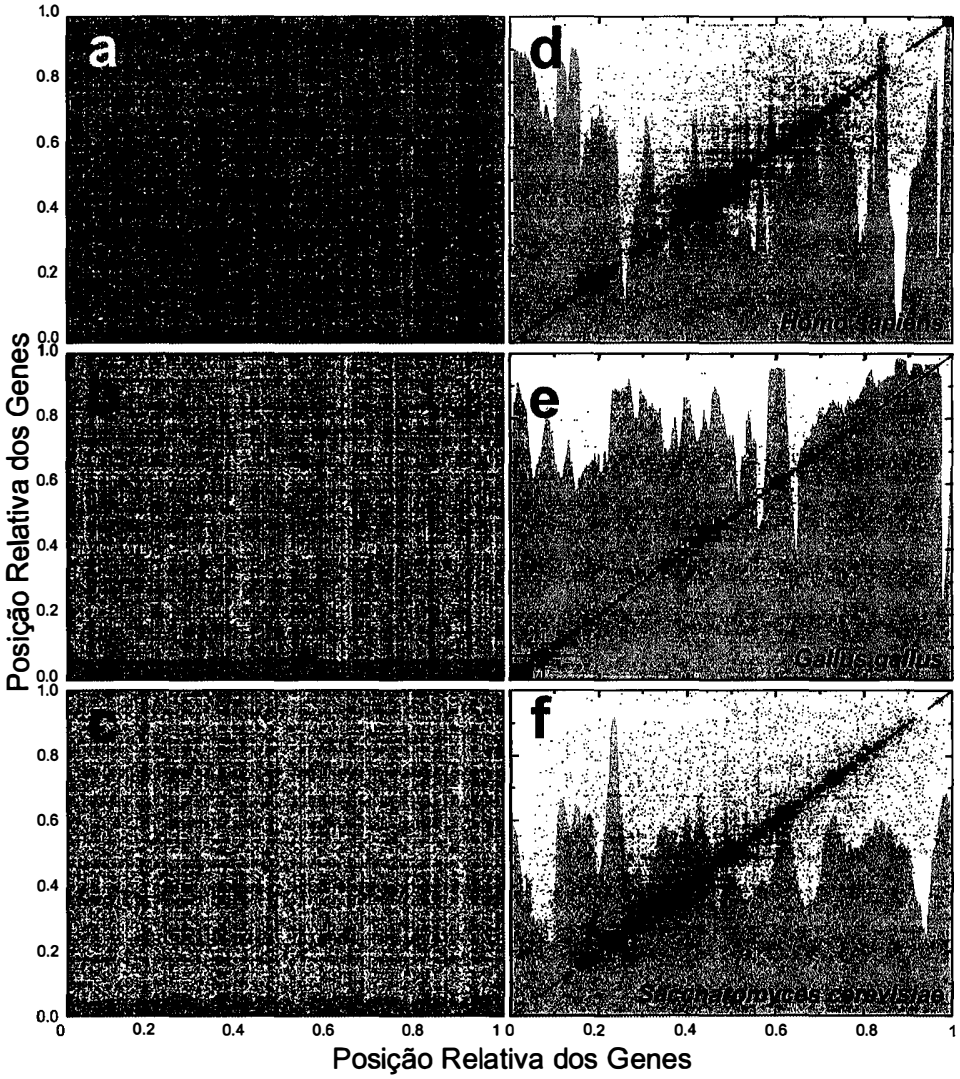


Figura 2:

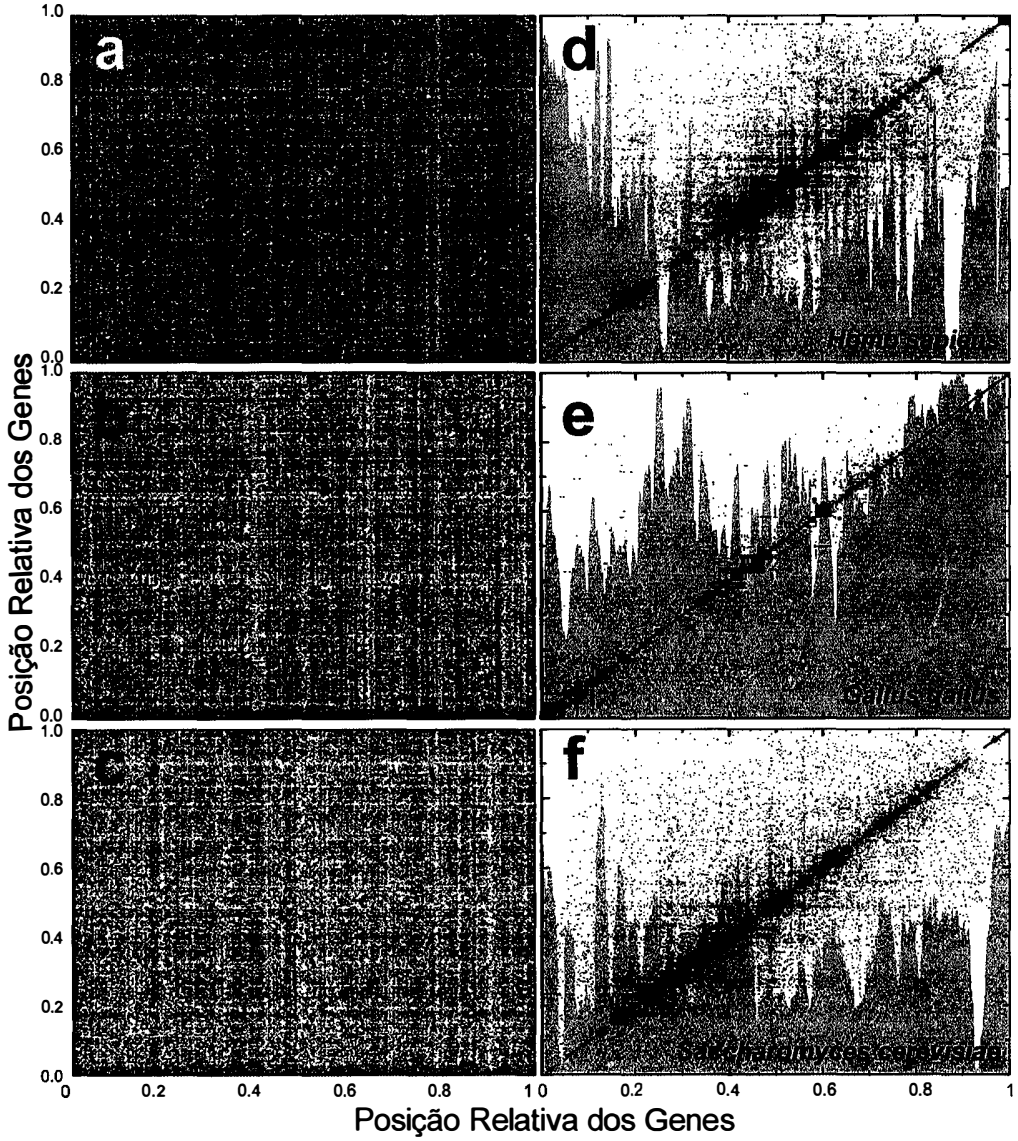

a
d
e
f
Posição Relativa dos Genes
Posição Relativa dos Genes
0.0
0.2
0.4
0.6
0.8
1.0

Figura 3:

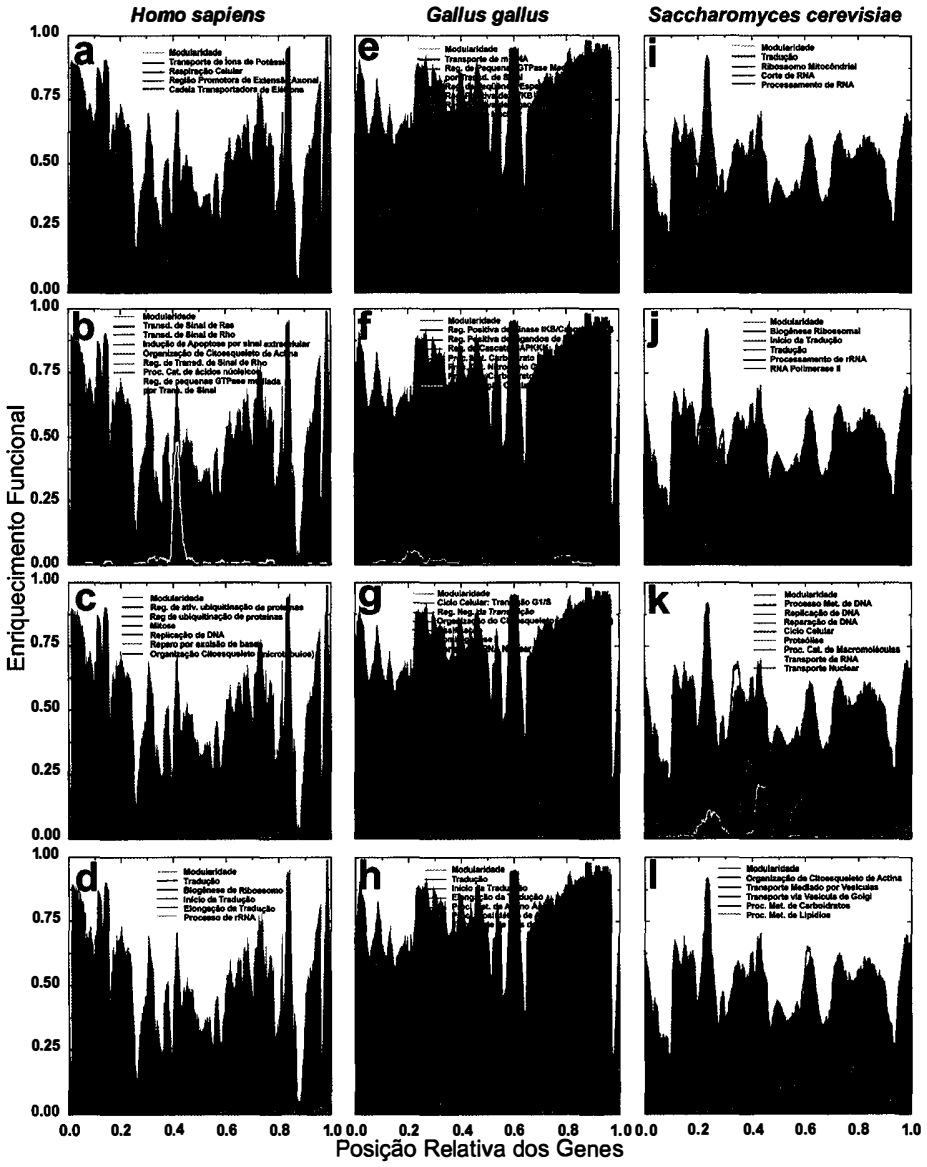

Homo sapiens
Gallus gallus
Saccharomyces cerevisiae
a
b
c
d
e
f
g
h
i
j
k
l
Modularidade
Transporte de Íons de Potássio
Respiração Celular
Região Promotora da Extensão Axonal
Cadeia Transportadora de Elétrons
Transd. de Sinal de Ras
Transd. de Sinal de Rho
Indução de Apoptose por sinal extracelular
Organização do Citoesqueleto de Actina
Reg. de Transd. de Sinal de Rho
Proc. Cat. de ácidos nucléicos
Reg. de pequenas GTPase mediada por Trans. de Sinal
Reg. de atv. ubiquitinação de proteínas
Reg de ubiquitinação de proteínas
Mitose
Replicação de DNA
Reparo por excisão de base
Organização Citoesqueleto (microtúbulos)
Tradução
Biogênese de Ribossomo
Início da Tradução
Elongação da Tradução
Processo de rRNA
Transporte de rRNA
Reg. Positiva de Cascata
Ciclo Celular: Transição G1/S
Reg. Neg. de Transcrição
Início da Tradução
Ribossomo Mitocôndrial
Corte de RNA
Processamento de RNA
Biogênese Ribossomal
Início da Tradução
Processamento de rRNA
RNA Polimerase II
Processo Met. de DNA
Replicação de DNA
Reparação de DNA
Ciclo Celular
Proteólise
Proc. Cat. de Macromoléculas
Transporte de RNA
Transporte Nuclear
Organização de Citoesqueleto de Actina
Transporte Mediado por Vesículas
Transporte via Vesícula de Golgi
Proc. Met. de Carboidratos
Proc. Met. de Lipídios
Enriquecimento Funcional
Posição Relativa dos Genes
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8
1.0

Figura 4:

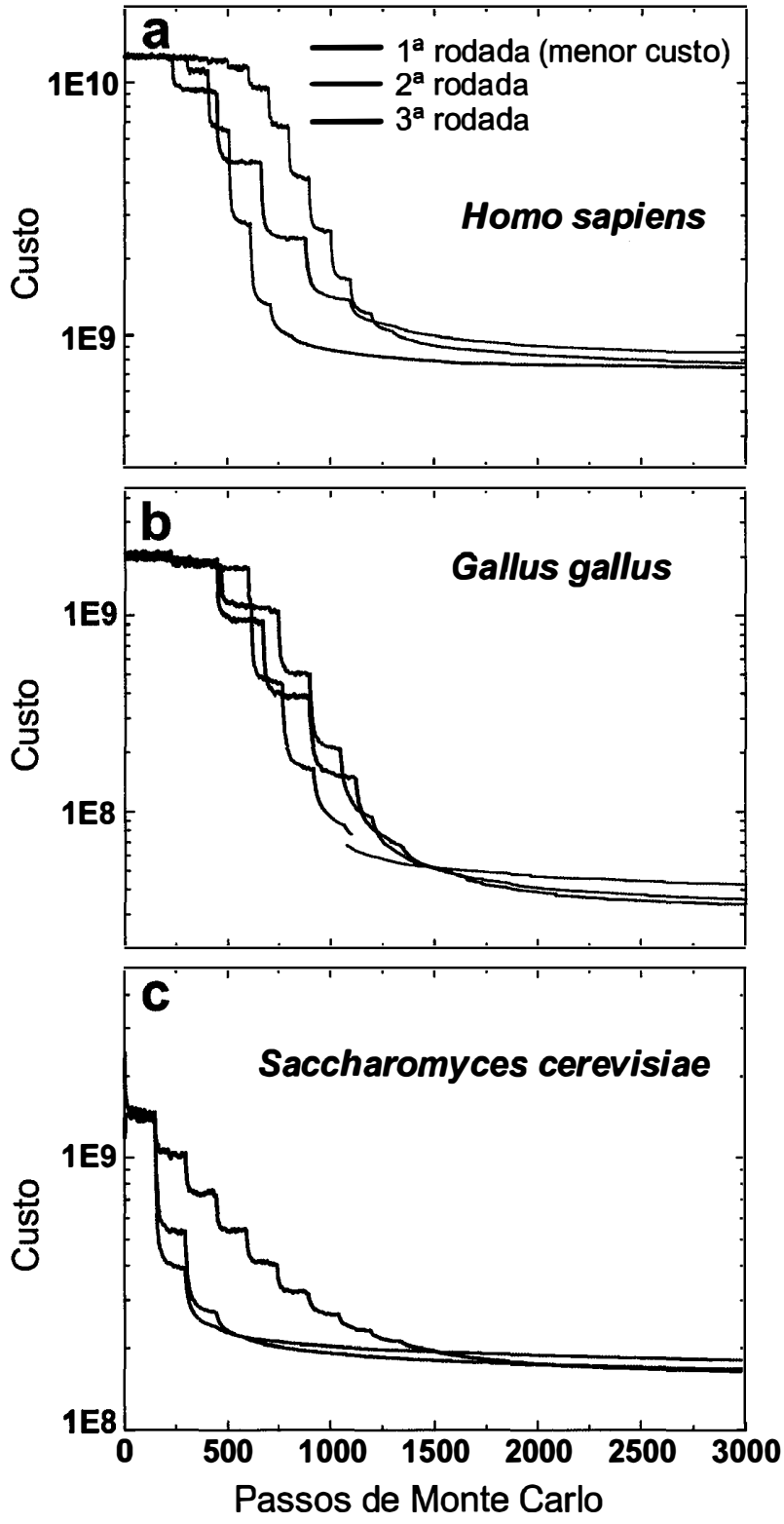

a
1ª rodada (menor custo)
2ª rodada
3ª rodada
Homo sapiens
1E10
1E9
Custo
b
Gallus gallus
1E9
1E8
Custo
c
Saccharomyces cerevisiae
1E9
1E8
Custo
0
500
1000
1500
2000
2500
3000
Passos de Monte Carlo

Figura 5:

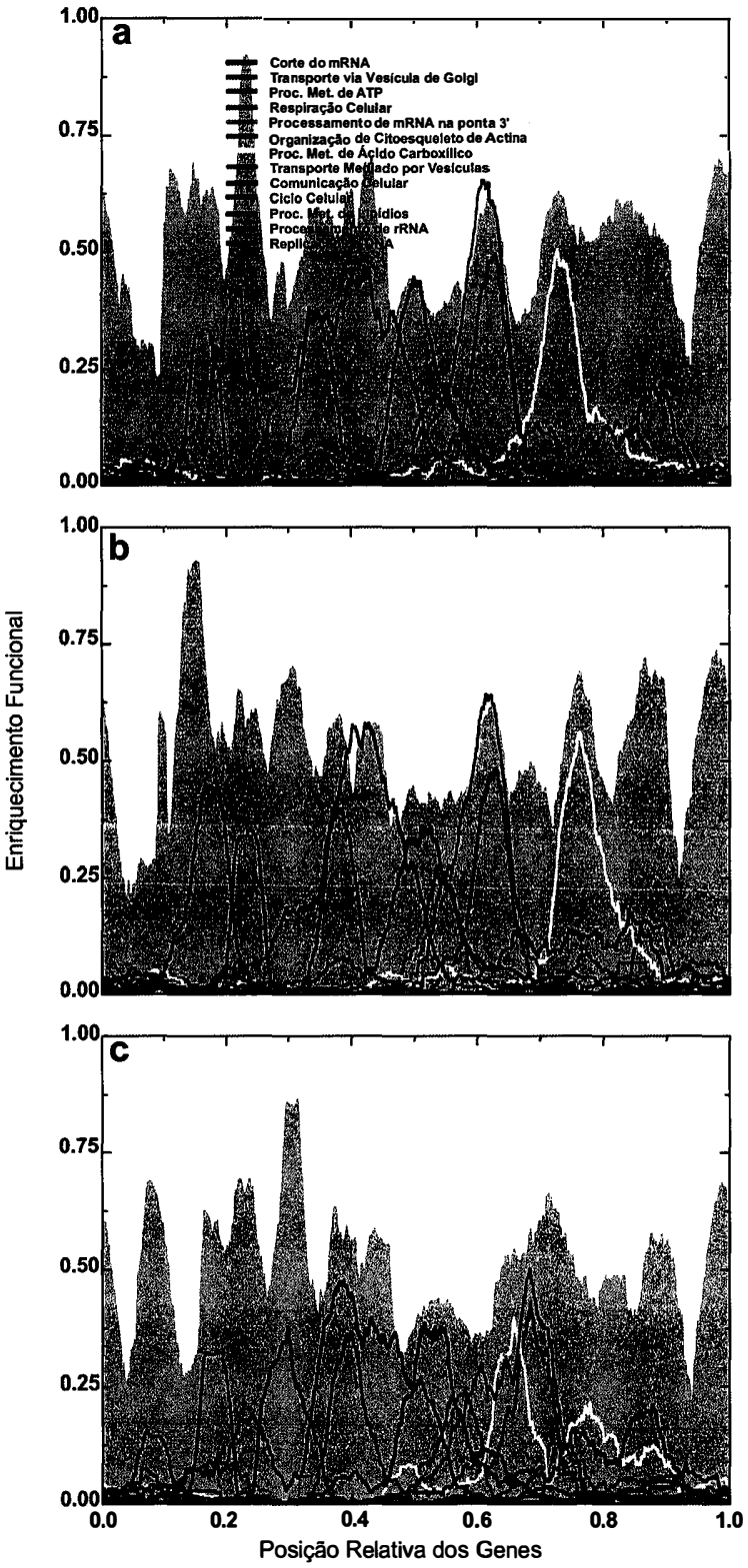

a
b
c
Corte do mRNA
Transporte via Vesícula de Golgi
Proc. Met. de ATP
Respiração Celular
Processamento de mRNA na ponta 3'
Organização de Citoesqueleto de Actina
Proc. Met. de Ácido Carboxílico
Transporte Mediado por Vesículas
Comunicação Celular
Ciclo Celular
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Enriquecimento Funcional
Posição Relativa dos Genes

Figura 6:

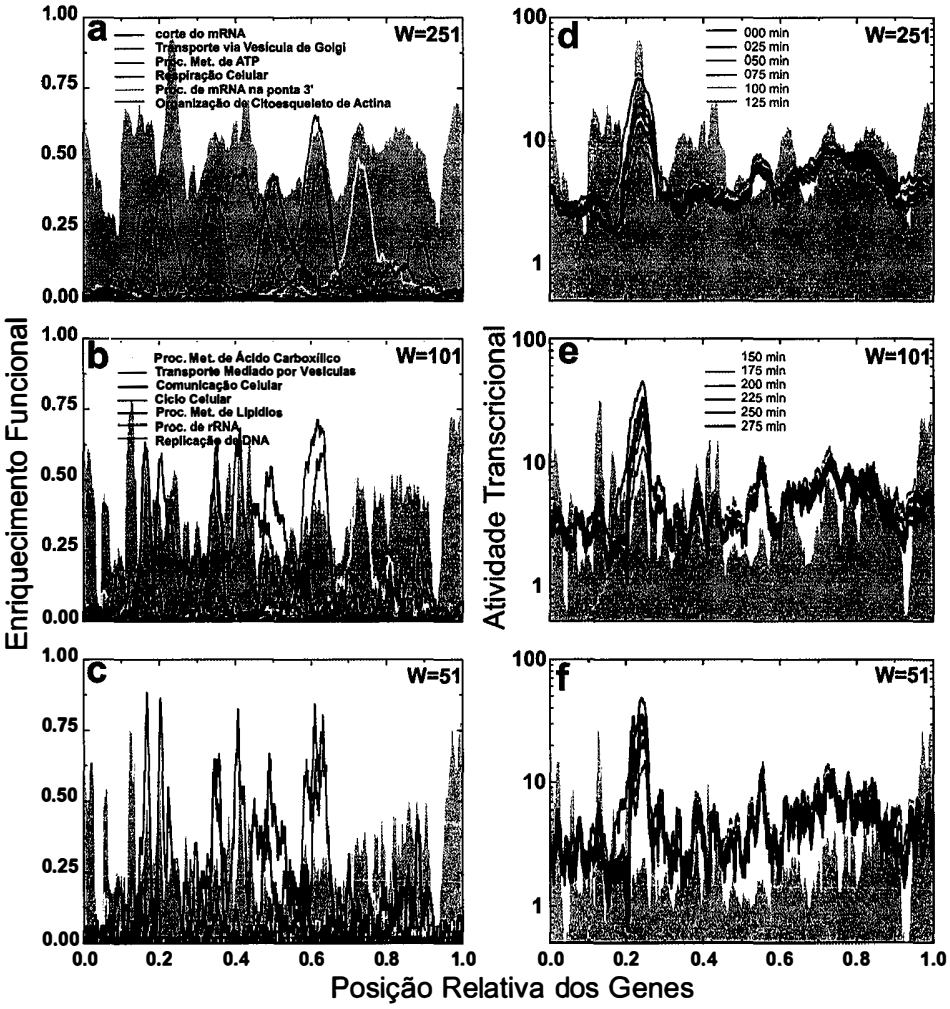

a
corte do mRNA
Transporte via Vesicula de Golgi
Proc. Met. de ATP
Respiração Celular
Proc. de mRNA na ponta 3'
Organização do Citoesqueleto de Actina
W=251
b
Proc. Met. de Ácido Carboxílico
Transporte Mediado por Vesículas
Comunicação Celular
Ciclo Celular
Proc. Met. de Lipídios
Proc. de rRNA
Replicação de DNA
W=101
c
W=51
d
000 min
025 min
050 min
075 min
100 min
125 min
W=251
e
150 min
175 min
200 min
225 min
250 min
275 min
W=101
f
W=51
Enriquecimento Funcional
Atividade Transcricional
Posição Relativa dos Genes

Figura 7:

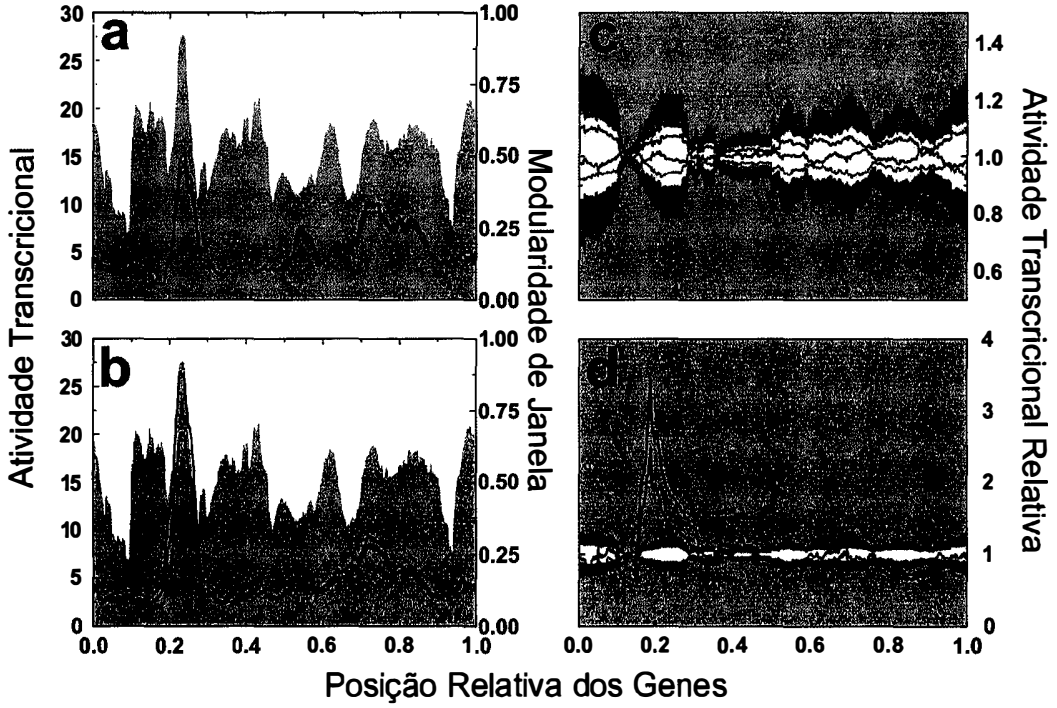


Figura 8:

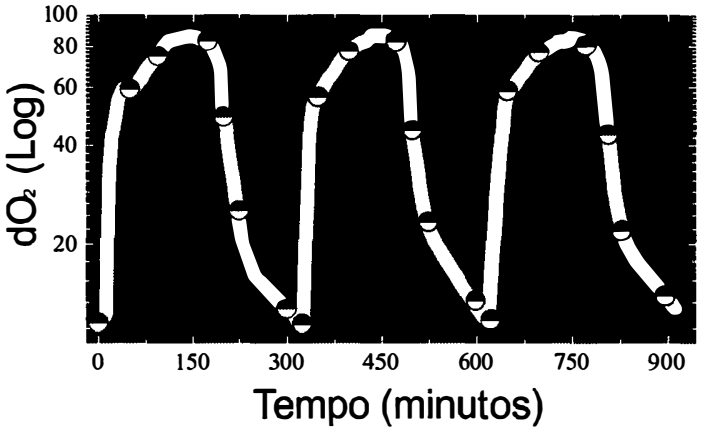


Figura 9:

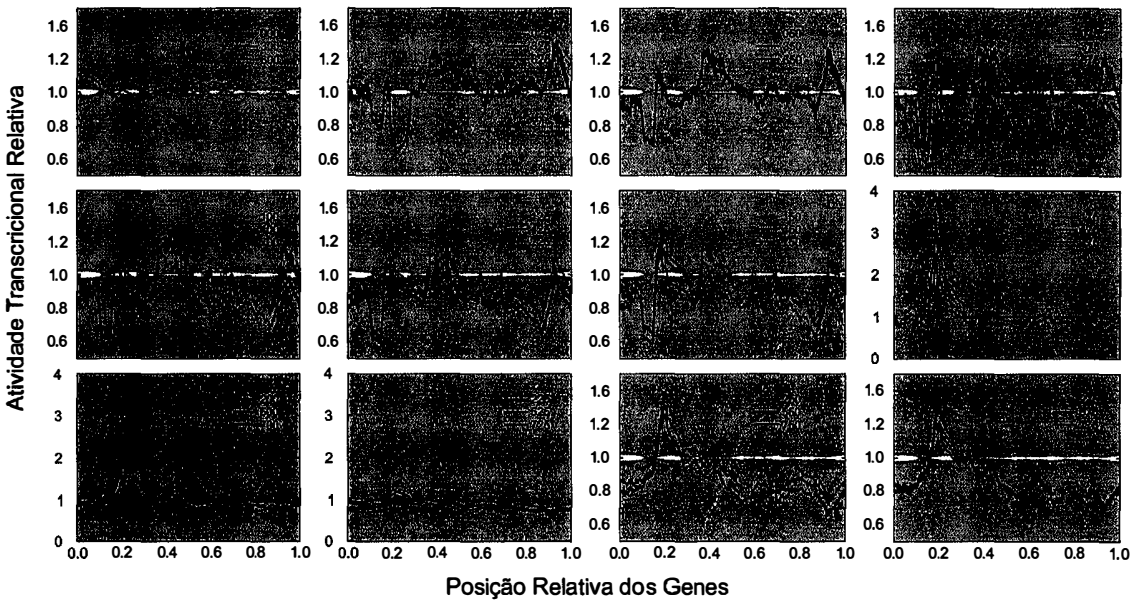


Figura 10:

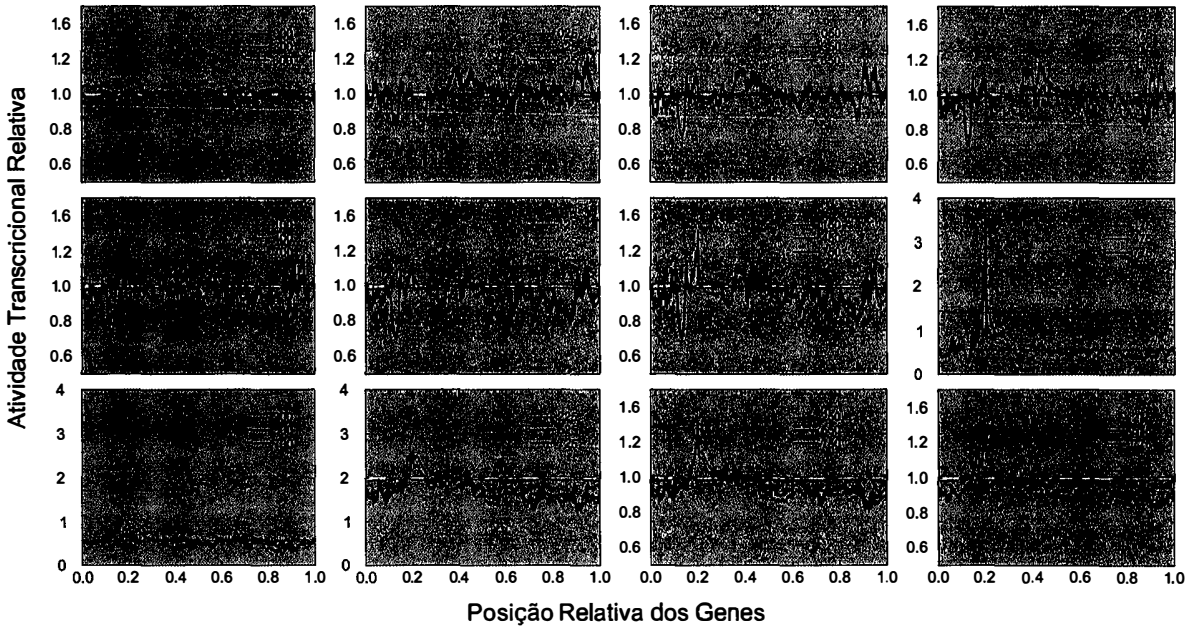
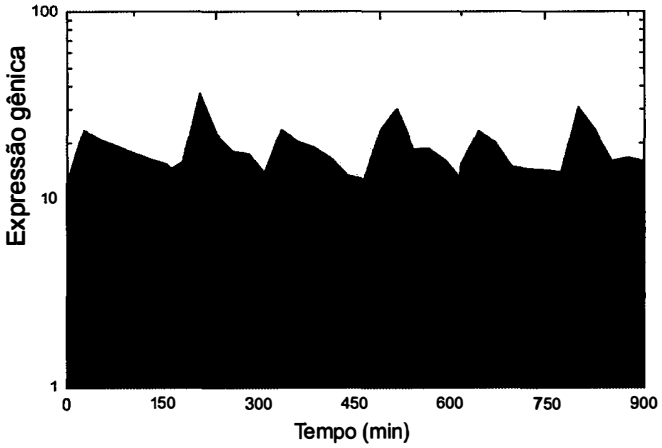


Figura 11:



Ox síntese de amino ácido, ribossomo
metabolismo de enxofre, metabolismo de RNA

R/C autofagia, peroxissomo, vacúolo
proteínas de choque térmico,
ubiquitina/proteassoma

R/B mitocôndria, replicação de DNA
histona, pólos do fuso

sinais de importação e exportação nuclear,
transporte de RNA, proteólise,
processo catabólico de macromoleculas

Figura 12:

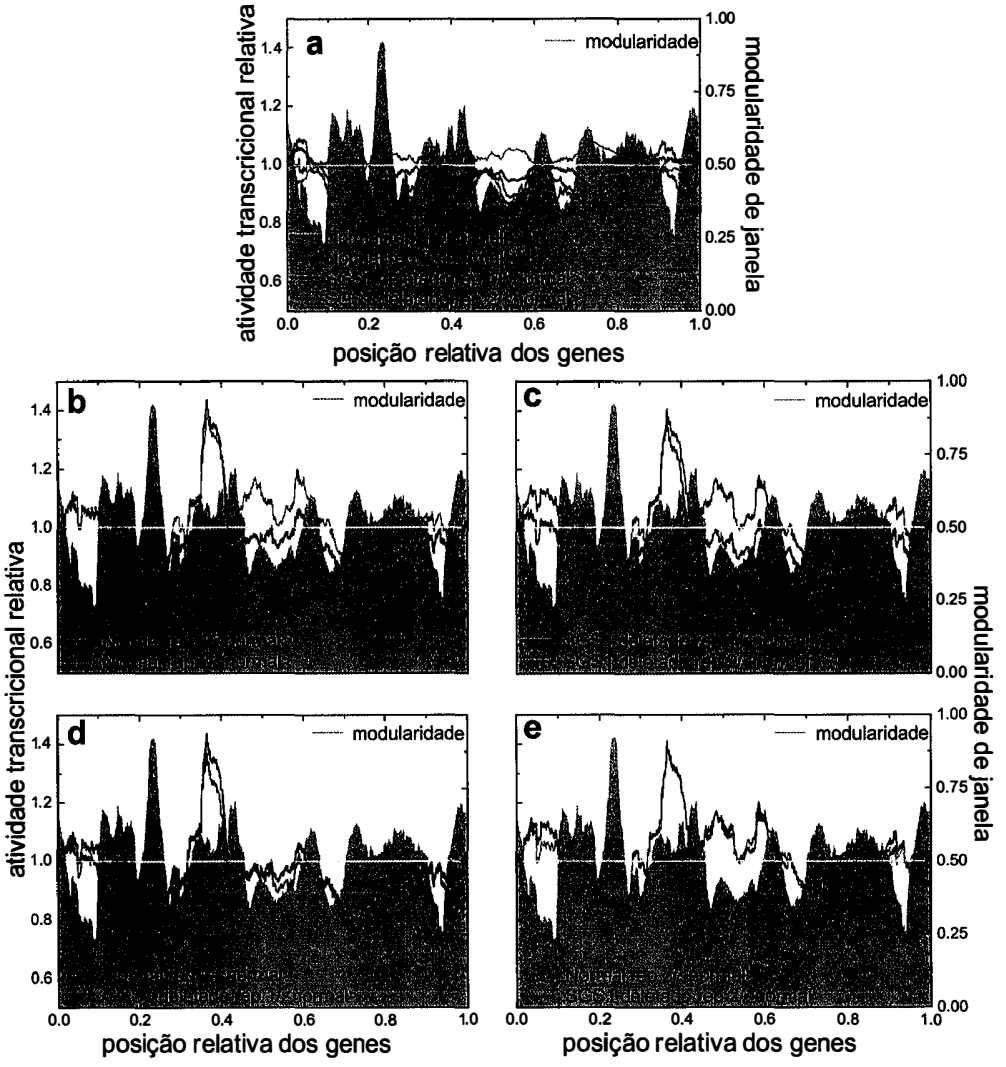


Figura 13:

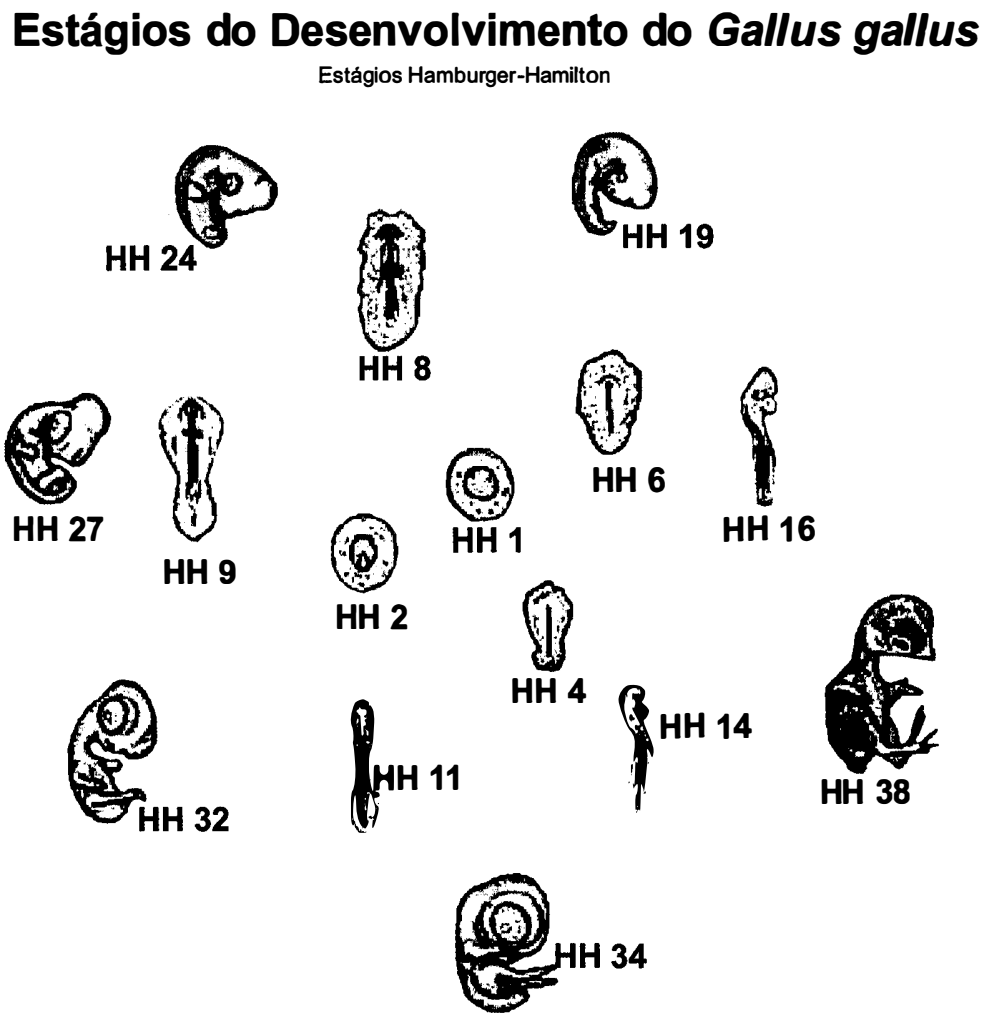

Estágios do Desenvolvimento do *Gallus gallus*
Estágios Hamburger-Hamilton
HH 24
HH 19
HH 8
HH 6
HH 16
HH 27
HH 9
HH 1
HH 2
HH 4
HH 14
HH 38
HH 32
HH 11
HH 34

Figura 14:

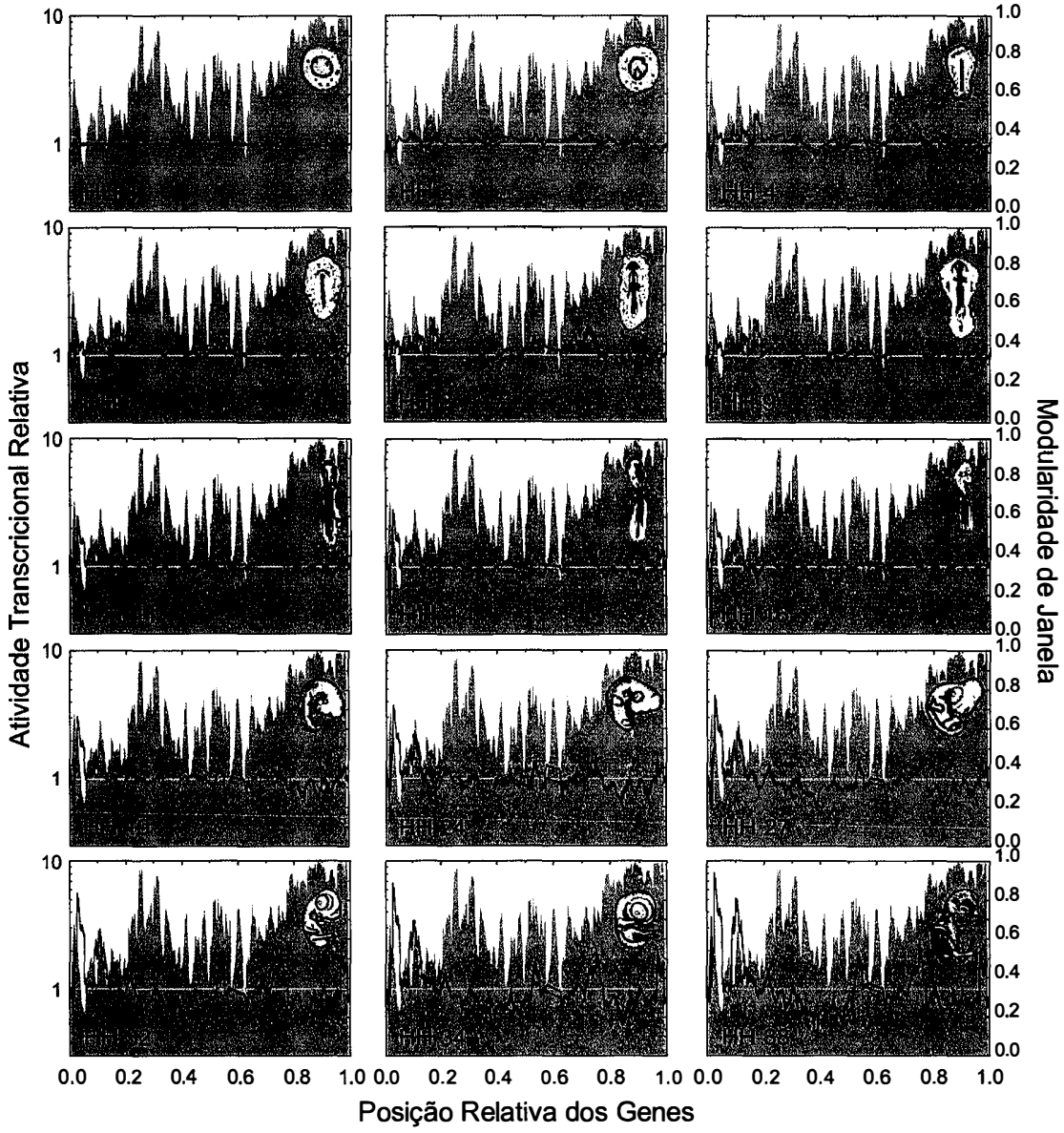

Atividade Transcricional Relativa
Modularidade de Janela
Posição Relativa dos Genes
10
1
0.0
0.2
0.4
0.6
0.8
1.0

Figura 15:

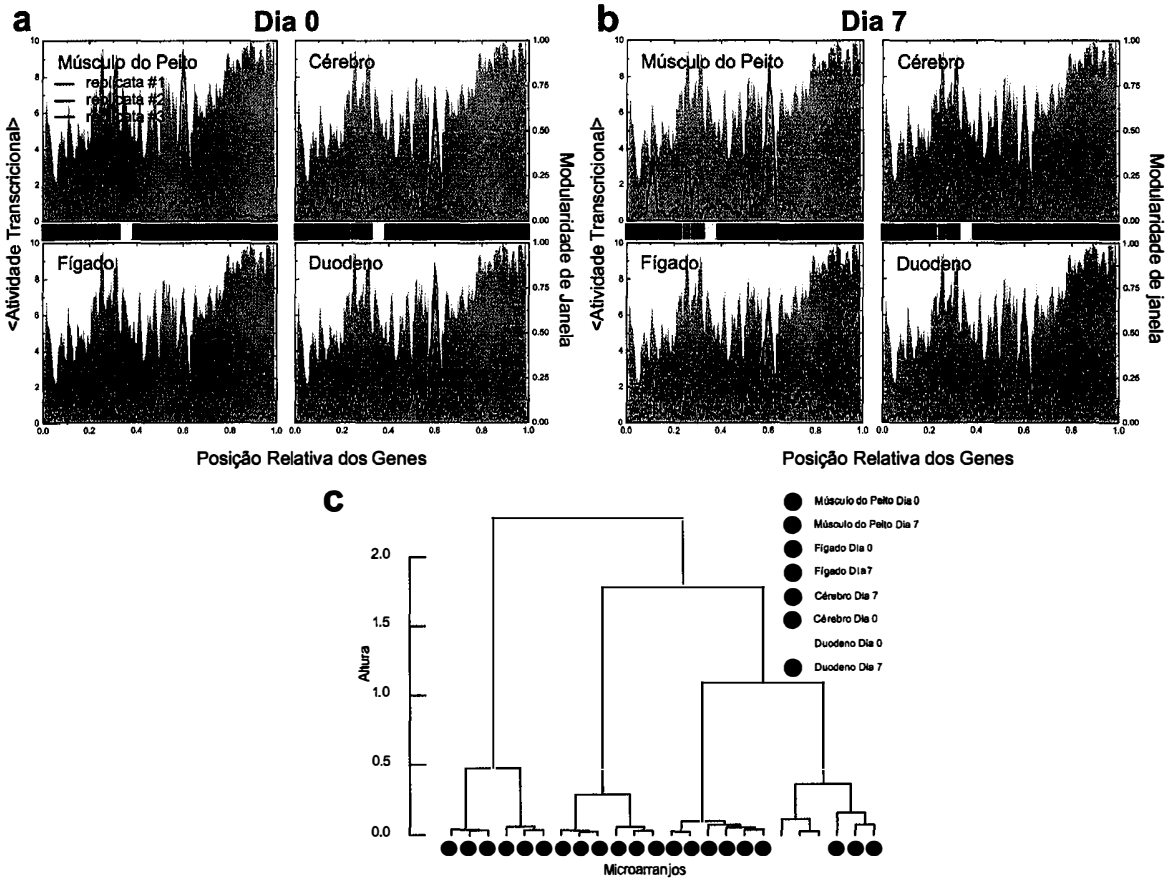

a
Dia 0
Músculo do Peito
Cérebro
Fígado
Duodeno
<Atividade Transcricional>
Modularidade de janela
Posição Relativa dos Genes
b
Dia 7
Músculo do Peito
Cérebro
Fígado
Duodeno
<Atividade Transcricional>
Modularidade de janela
Posição Relativa dos Genes
c
Músculo do Peito Dia 0
Músculo do Peito Dia 7
Fígado Dia 0
Fígado Dia 7
Cérebro Dia 7
Cérebro Dia 0
Duodeno Dia 0
Duodeno Dia 7
Altura
Microarranjos

Figura 16:

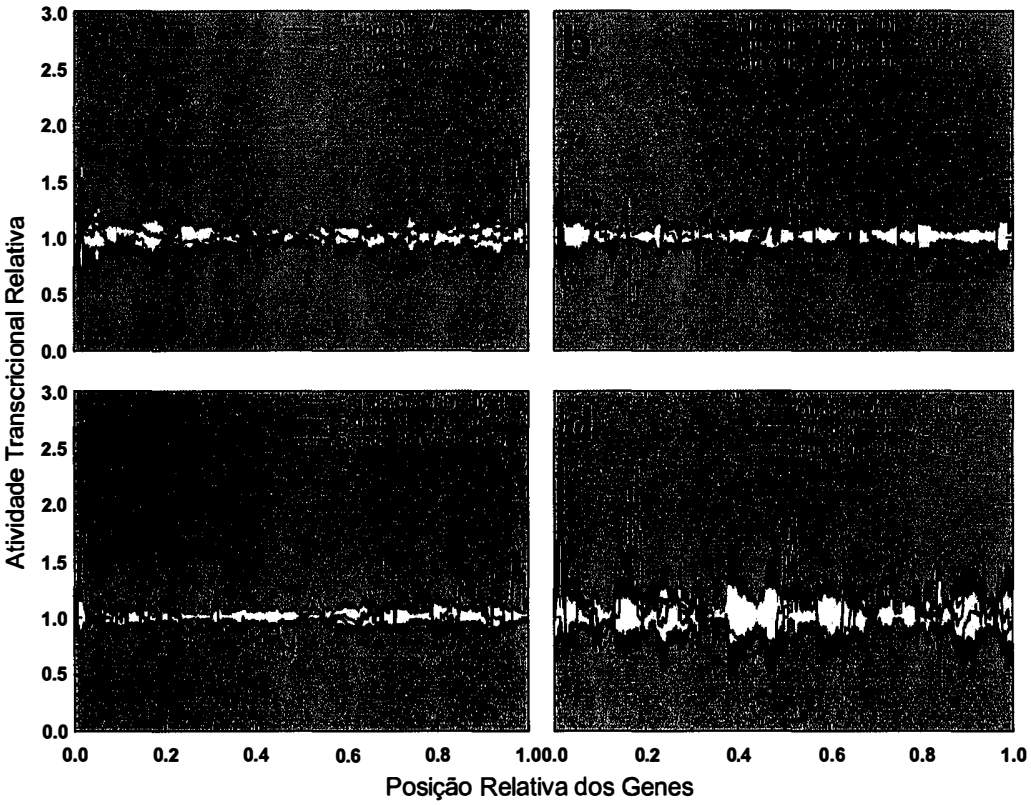

Atividade Transcricional Relativa
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Posição Relativa dos Genes

Figura 17:

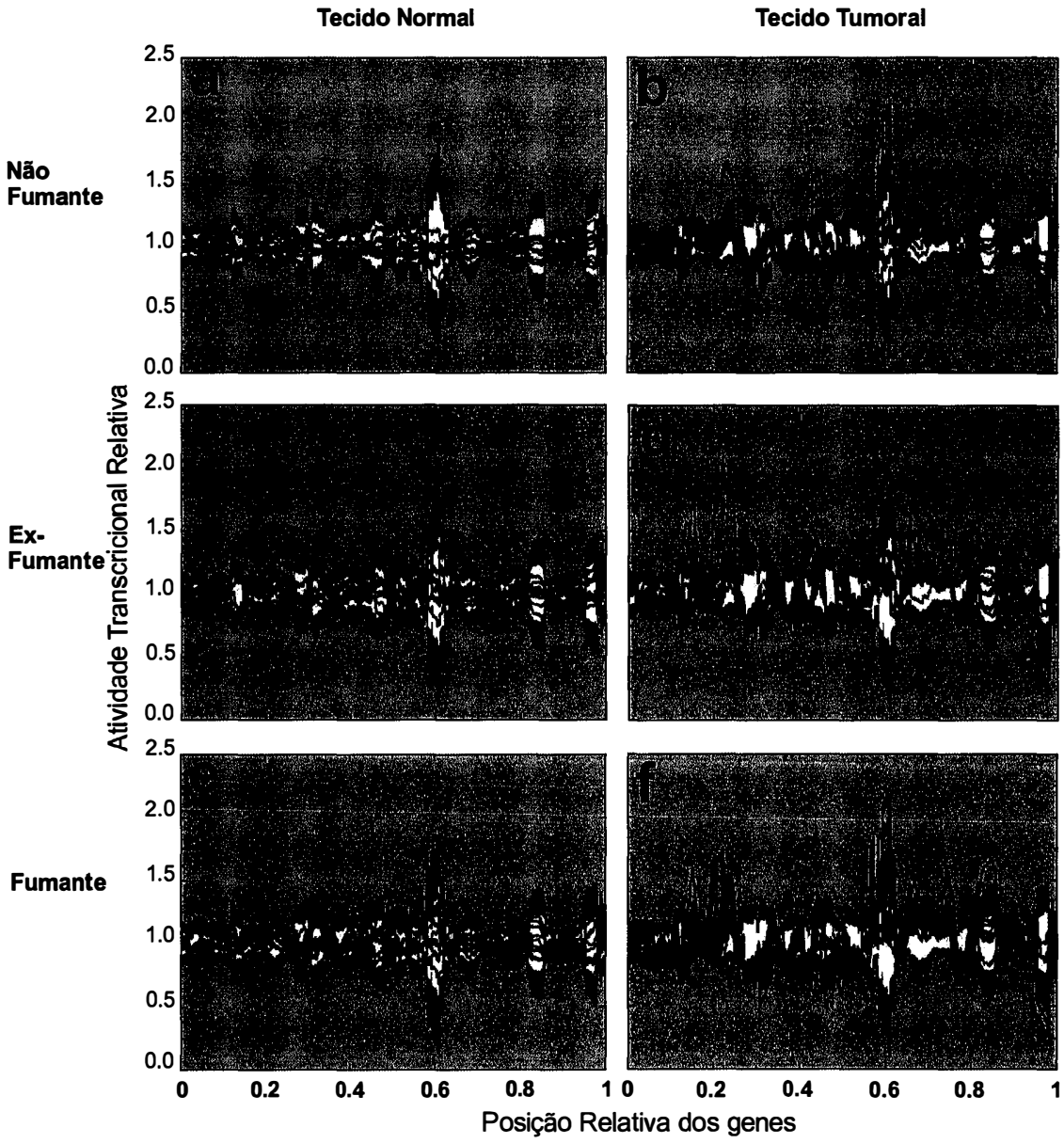
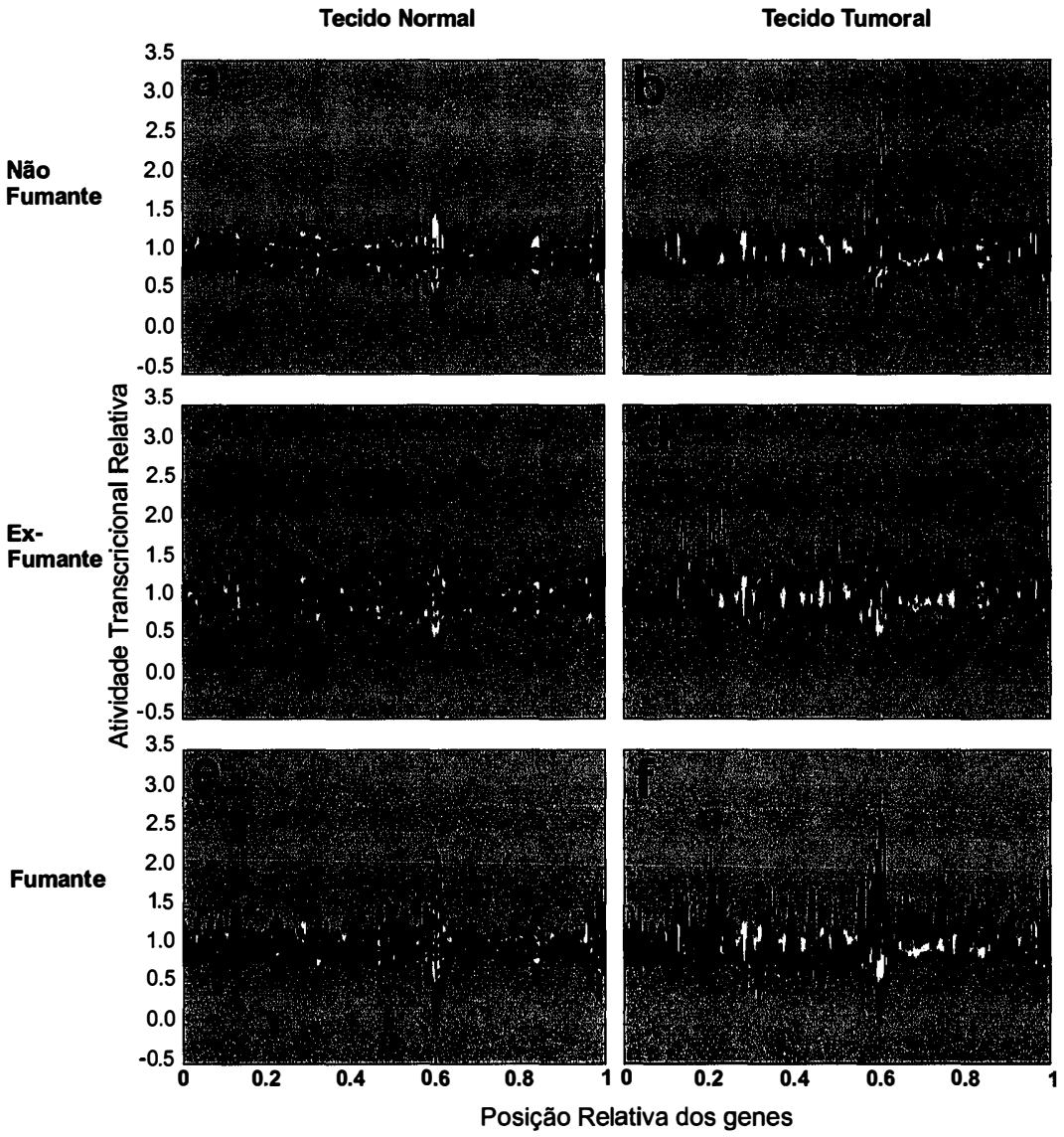

Tecido Normal
Tecido Tumoral
Não Fumante
Ex-Fumante
Fumante
a
b
c
d
e
f
Atividade Transcricional Relativa
2.5
2.0
1.5
1.0
0.5
0.0
0
0.2
0.4
0.6
0.8
1
Posição Relativa dos genes

Figura 18:


Tecido Normal
Tecido Tumoral
Não Fumante
Ex-Fumante
Fumante
Atividade Transcricional Relativa
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
0
0.2
0.4
0.6
0.8
1
Posição Relativa dos genes