UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

Alex Augusto Biazotti

Desenvolvimento de Ferramenta Computacional para integração de transcriptomas e redes biológicas: medidas de desempenho global

Botucatu
Fevereiro/2017

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

Alex Augusto Biazotti

Desenvolvimento de Ferramenta Computacional para integração de transcriptomas e redes biológicas: medidas de desempenho global

Dissertação apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento dos requisitos para a obtenção do título de Mestre no Programa de Pós-Graduação em Biotecnologia.

Área de Concentração: Biotecnologia
Orientador: Prof. Dr. José Luiz Rybarczyk Filho
Co-Orientadora: Prof.ª Dr.ª Agnes Alessandra Sekijima Takeda

Botucatu
Fevereiro/2017

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Biazotti, Alex Augusto.
Desenvolvimento de ferramenta computacional para integração de transcriptomas e redes biológicas : medidas de desempenho global / Alex Augusto Biazotti. - Botucatu, 2017

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk Filho
Coorientador: Agnes Alessandra Sekijima Takeda
Capes: 10302026

1. Transcriptoma. 2. Ontologia. 3. Análise de microarranjo. 4. Processamento eletrônico de dados.

Palavras-chave: Integração de dados; Microarranjo; Ontologia; Rede proteica; Transcriptograma.

Agradecimentos

- Ao CNPq por utilizarmos os recursos computacionais referentes aos processos 458810/2013-4 e 473789/2013-2
- Ao Professor Dr. José Luiz Rybarczyk Filho e Dra. Agnes Alessandra Sekijima Takeda pela excelente orientação, apoio e paciência;
- Ao meu pai José e a minha mãe Cleide pelo apoio e incentivo durante todo esse tempo;
- Aos meus amigos André Luiz Molan, Carlos Alberto de Oliveira Biagi Junior e José Rafael Pilan pela amizade e ajuda na confecção deste trabalho;
- Aos demais professores e funcionários que de alguma forma tenham contribuído para a realização deste trabalho;
- À Deus pela vida.

Resumo

A cada dia surgem novas tecnologias que possibilitam o estudo em larga escala dos RNAs transcritos por um organismo em condições específicas, com isso fornecendo uma grande quantidade de informações. No entanto as metodologias tradicionais não são capazes de análisar de forma eficiente esses dados por utilizar *cut-offs* pré-definidos, eliminando assim uma grande quantidade de genes não considerados diferencialmente expresso, e por consequência reduzindo a precisão e a acurácia do estudo. Esse trabalho propõe o aperfeiçoamento da metodologia do transcriptograma (modelo cruz), desenvolvido por Rybarczyk-Filho *et al.*, que realiza uma análise de forma global de um organismo, utilizando por sua vez redes proteícas e processos biológicos. Dentre as modificações realizadas estão: mudança no algoritmo de ordenamento para a redução do tempo de processamento da rede, adição de dois novos modelos "X" e "Anel", a automação dos processos de análise de dados de expressão gênica, enriquecimento funcional e da compilação de todas as informações em um gráfico. Para testar o aperfeiçoamento foram utilizadas duas séries de dados de expressão gênica, a GSE10072 e a GSE19804, referentes a amostras de câncer de pulmão. O modelo "Anel" apresentou a melhor redução do custo energético de uma matriz, aproximadamente 93%. Para a modularidade, o modelo "Anel" também teve o melhor desempenho. A automação dos processos de enriquecimento funcional, da análise dos dados de expressão e da compilação de todos os dados em forma gráfica diminui o tempo gasto para a aquisição e geração, além de aumentar a acurácia. Os resultados indicam que independentemente do hábito ou nacionalidade de um indivíduo, um mesmo tipo de câncer podem apresentar os mesmos conjuntos de processos biológicos alterados. A ferramenta não encontrou os mesmos processos biológicos indicados pelos *software* PAGE e GAGE, porém ele retornou processos mães ou filhos dos mesmos. A utilização desta ferramenta pode ser uma nova alternativa comparado aos demais métodos, devido a utilização de diversas informações adicionais ao conjunto de expressão gênica a ser analisado.

Abstract

Every day, new technologies are emerging that make it possible the large-scale study of RNAs transcribed by an organism under specific conditions, providing a huge amount of information. However, the traditional methodologies are not able to efficiently analyze these data due the use of pre-defined cut-offs, thus eliminating a large number of genes not considered differentially expressed, and consequently reducing precision andaccuracy of the study. This work proposes the improvement of the methodology of the Transcriptogram (model "Cross"), developed by Rybarczyk-Filho *et al.*, which performs an overall analysis of an organism, using protein networks and biological processes. Among the modifications made are: Modification in ordering algorithm to reduce the network processing time, addition of the two new "X" and "Ring" models, the automation of the processes of gene expression data analysis, functional enrichment and the compilation of all information in a graphic. To test the improvements, two sets of gene expression data were used, GSE10072 and GSE19804, corresponding to samples of lung cancer. The "Ring" model showed the best matrix energy cost reduction, approximately 93%. For modularity, the "Ring" model also had the best performance. The automation of functional enrichment processes, the analysis of expression data and the compilation of all data in a graphic form reduces the time spent for acquisition and generation, increasing the accuracy. The results indicate that regardless of habit of an individual, the same type of cancer may present the same sets of altered biological processes. The tool did not find the same biological processes indicated by the software PAGE and GAGE, but it returned their ancestor or child processes. The use of this tool may be a new alternative to the other methods, due the use of additional information to the set of gene expression to be analyzed.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

1.1 Microarranjo

No final do século XX, os pesquisadores tinham dificuldades de medir a expressão de vários genes de um organismo ao mesmo tempo, nesta época era possível medir apenas a expressão de poucos genes por vez. Mas com o passar dos anos, novas tecnologias foram desenvolvidas, uma dessas tecnologias foi a criação de um *chip* contendo várias sequências de nucleotídeos (cDNA ou oligonucleotídeo) denominado sonda. Com isso esse *chip* tornou-se uma ferramenta padrão para muitos laboratórios de pesquisa genômica(TSENG; GHOSH; FEINGOLD, 2012).

Um *chip* de microarranjo é composto por *spots/beads* e sondas. Os *spots/beads* são divisões no *chip* de microarranjo com identificadores que contém apenas parte de uma sequência com diversas cópias da mesma, denominada sonda, cada sonda é composta de 20-60 oligonucleotídeos (Figura 1.1), e ela é capaz de hibridizar com um fragmento de gene (Figura 1.2). A quantidade de *spot/bead* é diferente para cada organismo.

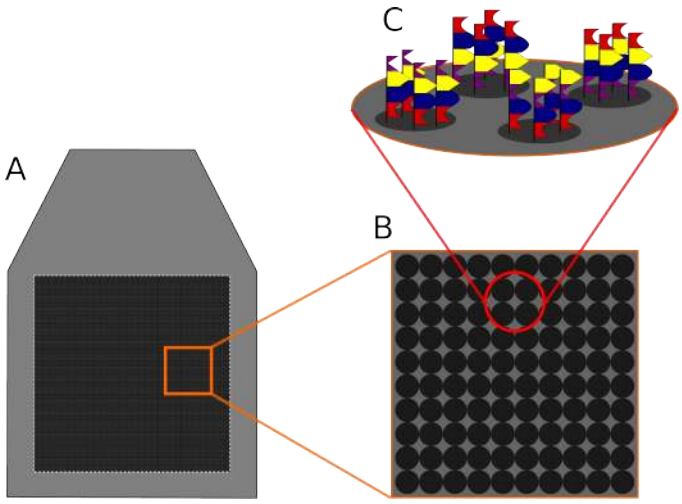

Figura 1.1: Representação dos componentes de um chip de microarranjo. (A) *chip* de microarranjo. (B) Pontos mais escuros, geralmente pretos, presentes no *chip* denominados *spots* ou *bead*. (C) Sondas presentes nos *spots*, onde cada sonda apresenta de 20-60 oligonucleotídeos.

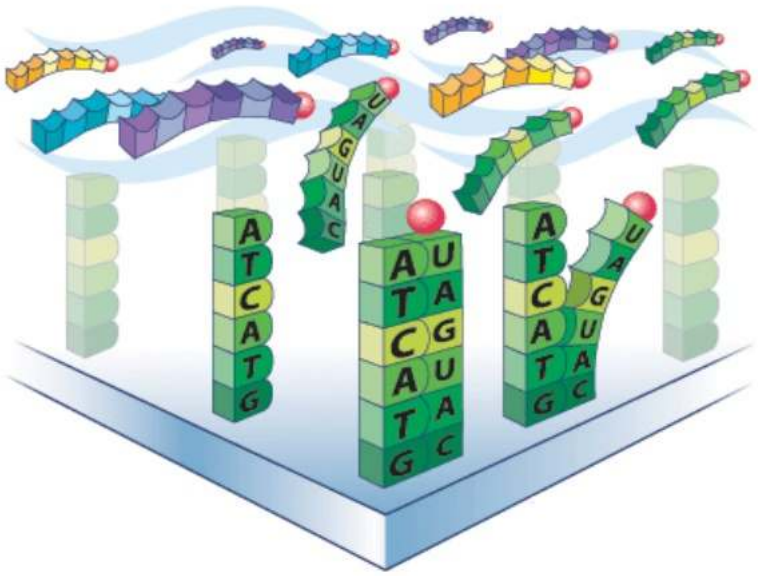

Figura 1.2: Hibridização da sonda do microarranjo com o fragmento do gene, representado pela sequência de nucleotídos com uma esfera vermelha. Os nucleotídeos presentes na sonda do chip irão combinar com os nucleotídeos presentes no fragmento dos genes. Adaptado de `www.essex.ac.uk/staff/langdon/genechip/`

A tecnologia mais utilizada realiza a hibridização de duas amostras (referência e teste). Com as sequências presentes no *chip* (LEUNG; CAVALIERI, 2003) e através de cálculos matemáticos e estatísticos obtêm-se os dados de expressão de cada sonda (XIE; PAN; KHODURSKY, 2005). Para a obtenção das informações de expressão é necessário duas amostras(referência e teste). Em seguida é realizada a extração dos mRNAs (RNA mensageiros) das amostras, então Aplica-se a enzima transcriptase reversa para obter o cDNAs, durante a obtenção dos cDNAs são utilizados nucleotídeos com os marcadores fluorescentes, o marcador vermelho para os mRNAs referentes a amostra do caso e verdes para a amostra referência. Com a obtenção dos cDNAs com os marcadores fluorescentes realiza-se a hibridização dos cDNAs com o *chip* de microarranjo, deixando eles agirem por algumas horas. Após a hibridização o chip é lavado para a remoção de cDNAs não hibridizados e colocado em um *scanner* que irá emitir um laser sobre o *chip*, essa emissão realizada duas vezes, onde uma vez irá emitir na frequência para captar a tonalidade vermelha e depois na frequência para captar a tonalidade verde. Depois da captação, as mesmas são mescladas através de um algoritmo estatístico, onde apresenta novas colorações como tons que variam de amarelo até laranja, essa nova variação a representação que houve a hibridização tanto da amostra teste quanto da amostra referência naquele *spot/bend*. Entretanto ele pode apresentar a coloração preta que é referente a não expressão de nenhuma das amostras (Figura 1.3).

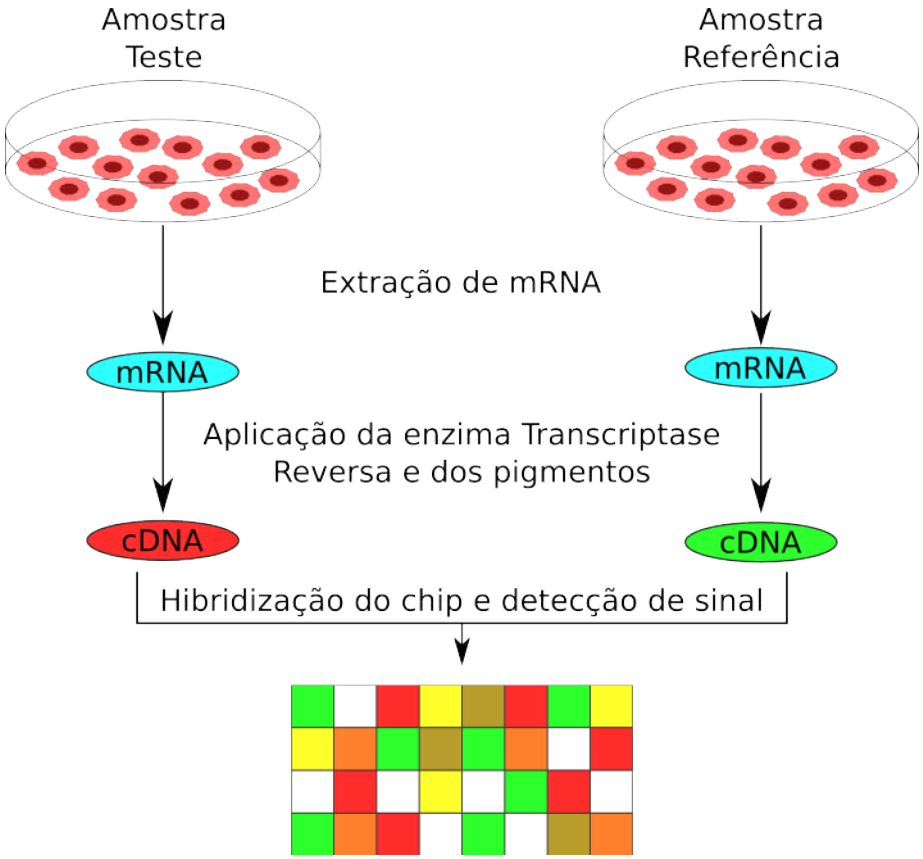

Figura 1.3: Representação referente a extração da informação da expressão dos genes, extração do mRNA das amostras e aplicação da enzima transcriptase reversa juntamente com nucleotídeos com marcadores fluorescentes verde e vermelho. Por fim são combinados e aplicados no chip, deixando hibridizar por algum tempo e inseridos em um sistema computadorizado para a extração da luminosidade dos genes hibridizados.

Com a aquisição dos níveis de expressão registrados pelo *chip*, através da frequência luminosa, é necessária a extração dos dados através de uma técnica de normalização para que seja possível a manipulação/estudo da expressão dos genes. Existem diversas técnicas de normalização, mas as mais utilizadas são a *Affymetrix Microarray Suite 5.0* (MAS5), *Robust Multi-array Analysis* (RMA) e *Robust Multi-array Analysis with correction for GC content* (GCRMA). Por meio destas técnicas torna-se possível o estudo da expressão dos genes por meio de cálculos como *fold-change*, expressão média, teste-t, p-valor para cada gene (DALMAN et al., 2012), e com o resultados desses cálculos, os pesquisadores são capazes de aplicar alguns critérios nos valores para verficação de quais genes estão superexpressos ou subexpressos.

1.2 Tecnologias de Microarranjo

Apesar do método para a utilização do chip ser muito semelhante entre as empresas fabricantes, cada *chip* apresenta uma característica única referente a empresa, podendo ser o número de oligonucleotídeos presentes em uma sonda até na forma de confecção do *chip*. A seguir são apresentadas três empresas fabricantes dos *chips* e das tecnologias de microarranjo.

1.2.1 Affymetrix

A Affymetrix foi a primeira empresa a trabalhar com *chips* de microarranjo destinado ao comércio (LOCKHART et al., 1996). Seu *chip* conta com 11-20 sondas em cada *spot*, cada sonda apresenta 25 nucleotídeos *perfect match* (PM) (BARNES, 2005). Na maior parte de seus arranjos, a Affymetrix dispõe para cada PM um *mismatch* (MM) (Figura 1.4), a partir 13º base ele difere para mensurar hibridizações não-específicas (GAUTIER et al., 2004; IRIZARRY et al., 2003a).

T	T	A	C	C	C	A	G	C	T	T	C	C	T	G	A	G	G	A	T	A	C	Sequência Teste
A	A	T	G	G	G	T	C	G	A	A	G	G	A	C	T	C	C	T	A	T	G	Sonda *Perfect Match*
A	A	T	G	G	G	T	C	G	A	A	C	G	A	C	T	C	C	T	A	T	G	*Sonda Mismatch*

Figura 1.4: Representação das sondas *Perfect Match*(PM) e *Mismatch* (MM). A sequência em ciano é uma sequência a ser hibridizada com as sondas, onde as sondas PM apresentam todos os nucleotídeos complementares a essa sonda, sendo que onde na estiver na sequência teste timina (T), adenina (A), citosina (C) e guanina (G) na sonda estará, respectivamente, adenina (A), timina (T), guanina (G) e citosina (C). Já as sondas MM apresentará ao menos um nucleotídeo não complementar a sequência teste, como representado em vermelho, onde ele apresenta uma citosina (C) no lugar de uma guanina (G).

1.2.2 Agilent

Assim como a Affymetrix, a Agilent também utiliza os *spots*, entretanto difere-se na preparação de seu *chip*. Ele é fabricado com a tecnologia de *SurePrint ink-jet* (Figura 1.5), que é uma tecnologia desenvolvida pela Agilent (GERSHON, 2002). A composição do *chip* é muito semelhante ao da Affymetrix, contendo de 11-20 sondas, porém o tamanho de cada sonda possuí 60 nucleotídeos (ZAHURAK et al., 2007).

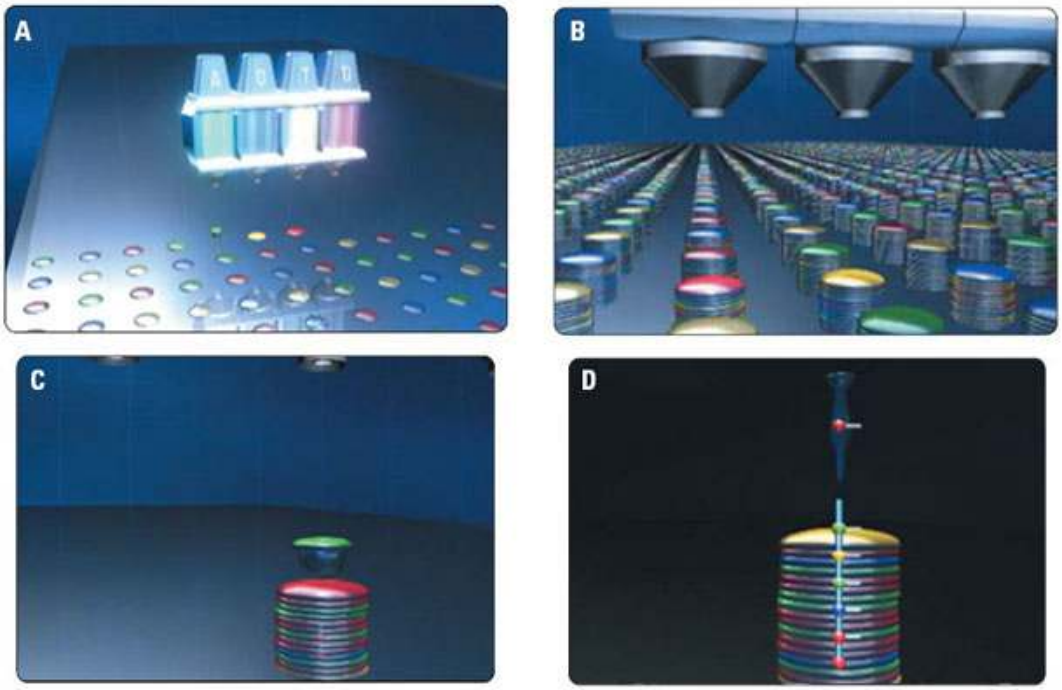

Figura 1.5: Representação do funcionamento da tecnologia *SurePrint ink-jet*. **A**. A primeira camada de nucleotídios é depositada na superficie do *chip*. **B**. Novas camadas são, precisamente, impressos sobre as antigas camadas. **C**. Aproximação de uma cadeia onde está sendo inserido uma nova camada, sendo complementada com a imagem **D**. que representa a disposição da sequência de nucleotídeos dentro da cadeia e a adição de um novo nucleotídeo/camada. Retirado de `https://www.genomics.agilent.com/`

1.2.3 Illumina

Diferente da tecnologia das outras empresas, a Illumina possui a tecnologia de *BeadArray*, que é baseada em uma montagem aleatória das sondas em cada *bead* (FAN et al., 2006). A quantidade de sondas presentes em cada *bead* é de aproximadamente 30 cópias (BARNES, 2005), tornando o pré-processamento e o controle de qualidade eficiente e significativamente diferente das outras tecnologias (DU; KIBBE; LIN, 2008).

A Illumina apresenta duas plataformas de microarranjo (FAN et al., 2006; STEEMERS, 2005) a *Sentrix Array Matrix* que contém uma configuração de 96 análises simultâneas com aproximadamente 1.500 variedades de *beads* por análise, e a *Sentrix BeadChip* contendo uma configuração de 1 a 16 análises simultâneas com aproximadamente 24.000 variedades de *bead* (DUNNING et al., 2007).

1.3 Normalização

A normalização, como citado anteriormente, é uma técnica essencial para a extração das expressões obtidas através das diferentes intensidades luminosas das amostras biológicas, identificando assim genes diferencialmente expressos (YANG; THORNE, 2003). Existem diversos métodos para a realização da normalização, cada qual com a sua característica, porém as mais utilizadas pela comunidade científica são: *MicroArray Suit 5*, *Robust Multi-Array Average* e *GC Robust Multi-Array Average*.

1.3.1 *MicroArray Suite 5* (MAS5)

A técnica de normalização MAS5 foi desenvolvida pela Affymetrix, e utiliza como parâmetros para a correção do *background* PM e MM, tornando assim os resultados de baixa expressão diferencial mais vísivel (PEPPER et al., 2007). Isso é possível através do cálculo estatístico realizado por esse método que realiza a subtração dos valores de MM de cada PM par (IRIZARRY, 2003). A expressão dos genes não é apresentada na forma de base logarítmica como os demais métodos de normalização, além disso, a técnica informa o p-valor e o "*detection call*". O *detection call* informa se o transcrito está "presente", "ausente" ou "marginal", ou seja, se o resultado normalizado do transcrito é "significante", "não-significante" ou não se altera em ambos os casos (referência e teste) (PEPPER et al., 2007).

1.3.2 *Robust Multi-Array Average* (RMA)

A metodologia RMA é uma das mais utilizadas para a normalização de microarranjos (ZIMMERMANN; LESER, 2010). Ao contrário do MAS5, esta baseia-se apenas nos PM de cada microarranjo, com isso, ao realizar a correção do *background* seu resultado nunca será negativo. Após a correção do *background* os valores são transformado para $\log_2$ e então aplicado o cálculo estatístico para normalizar os dados (IRIZARRY et al., 2003b). Por não utilizar os MM nos cálculos, essa metodologia garante que haverá menos ruído no resultado.

1.3.3 *GC Robust Multi-Array Average*(GCRMA)

A metodologia GCRMA é uma modificação da RMA, em seu algoritmo foi incorporado a sequência da sonda no ajuste do *background* (GHARAIBEH; FODOR; GIBAS, 2008). Ela utiliza o modelo de Naef e Magnasco (NAEF; MAGNASCO, 2003). Ao realizar essa etapa extra

em relação ao RMA, o GCRMA consegue detectar com melhor eficácia genes diferencialmente expressos que tiveram uma baixa intensidade de detecção (WU, 2009).

1.4 Problemas nas análises

As metodologias tradicionais de análise de genes diferencialmente expressos utilizam um valor fixo de *cut-off*, geralmente *two-fold change* para classificá-los em superexpresso (*up-regulated*) ou subexpresso (*down-regulated*) (LEUNG; CAVALIERI, 2003). Ao aplicar este *cut-off*, é possível descartar uma gama de genes que podem possuir um *fold-change* inferior mas com significância estatística (GUSNANTO; CALZA; PAWITAN, 2007; KIM et al., 2002). O *cut-off* não é um teste estatístico, apenas uma comparação direta entre os logarítmos de cada sonda em relação a uma referência, com isso a quantidade de falsos-positivos e falsos-negativos aumentam drasticamente (CUI; CHURCHILL, 2003). Uma forma de contornar esse problema é a utilização de métodos mais complexos envolvendo não somente os valores de expressão dos genes, mas combinando redes biológicas e ontologias para avaliar a expressão.

1.5 Redes

Redes, de uma forma simplificada, compreende um conjunto de elementos que interagem entre si, como redes sociais, redes tecnológicas, redes biológicas, e muitas outras. No caso de redes biológicas, podemos ter uma rede de interação de proteínas, onde cada proteína receberá o nome de vértice (nós/nodos) e as ligações entre as proteínas são chamadas de arestas (ligações/interações).

Além disso, as redes podem apresentar uma nova característica quanto ao seu direcionamento, podendo serem chamadas de redes direcionadas ou redes não-direcionadas. As redes direcionadas apresentam a direção, ou seja, um elemento “a” interage com um elemento “b”, porém o elemento “b” não interage de volta com o elemento “a” (Figura 1.6 A). Já as redes não-direcionadas não apresentam uma direção, ou seja, um elemento “a” interage com um elemento “b”, e o elemento “b” interage de volta com o elemento “a” (Figura 1.6 B).

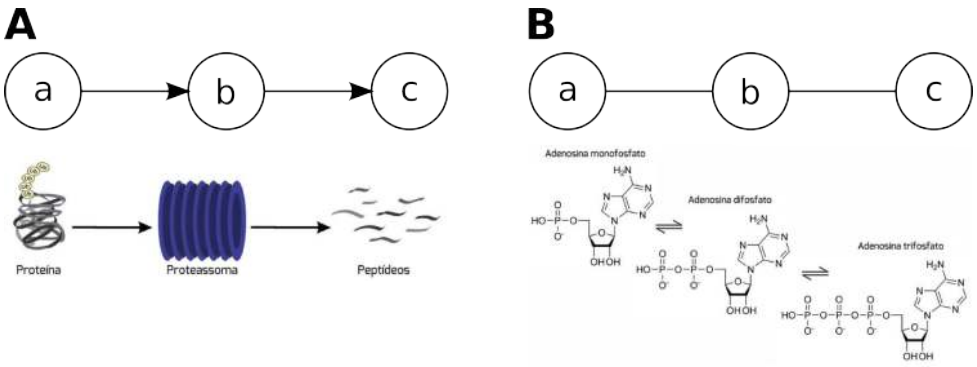

Figura 1.6: Exemplos de rede direcionada e não-direcionada. (A) representação gráfica de uma rede direcionada e a via de degradação de proteína que foi baseada; (B) representação gráfica de uma rede não-direcionada e a reação reversa de fosforilação de adenosina difosfato. Adaptado de (VERLI et al., 2014).

Uma rede de N vértices pode ser completamente transformada em uma matriz de adjacência A, onde A_{ij} assume valor 1 quando houver uma interação entre os elementos, e assume o valor 0 quando não houver qualquer interação entre os elementos. Ao realizar essa transformação em uma rede tem-se como resultado uma matriz. Ao transformar uma rede direcionada em matriz é obtido uma matriz com uma grande quantidade de valores 0 (Figura 1.7), entretanto ao transformar uma rede não-direcionada, temos uma matriz simétrica (Figura 1.8). Para se calcular as centralidades, prefere-se utilizar redes não-direcionas, mesmo que exista uma maior quantidade de interações, a quantidade de cálculos necessários é bem menor, pois como a matriz é simétrica, então é possível considerar apenas a diagonal superior ou inferior para realizar os cálculos.

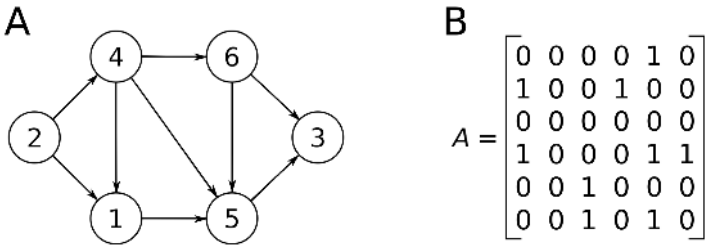

Figura 1.7: Transformação de uma rede direcionada em uma matriz de adjacência. (A) Rede de interação, onde os valores representam os vértices da matriz; (B) Matriz de adjacência, onde foi tomado como base a rede de interação anterior.

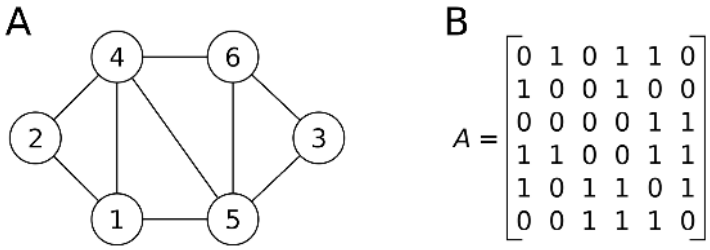

Figura 1.8: Transformação de uma rede não-direcionada em uma matriz de adjacência. (A) Rede de interação, onde os valores representam os vértices da matriz; (B) Matriz de adjacência, onde foi tomado como base a rede de interação anterior.

1.5.1 Centralidades

1.5.1.1 Conectividade

A conectividade (ou grau do nó) é a representação de interações ou vizinhos que um vértice possui, apresentado pela equação 1.1:

$$k_i = \sum_{j}^{N} a_{ij}, \tag{1.1}$$

onde N representa o total de vértices, a_{ij} são os elementos da matriz de adjacência A e k_i a conectividade do vértice. Como exemplo, temos a Figura 1.9.

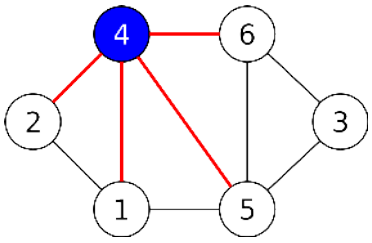

Figura 1.9: O vértice 4 na rede apresenta 4 ligações com os seus respectivos vizinhos, logo a sua conectividade é 4.

A partir deste cálculo podemos determinar elementos que são altamente conectados, chamados de *hubs*. Além de definir os *hubs*, é possível determinar a conectividade média da rede $\bar{k}$ (Equação 1.2).

$$\bar{k} = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{1.2}$$

Com a conectividade também é possível verificar a distribuição de conectividades de uma rede, ou seja, verifica a fração de vértices que possuí conectividade k.

1.5.1.2 Coeficiente de clusterização

O coeficiente de clusterização (C_i) é a representação da interação entre os vizinhos do vértice i, dado pela Equação 1.3:

$$C_i = \frac{2N_i}{k_i(k_i-1)} = \frac{1}{k_i(k_i-1)} \sum_{j=1}^{N} a_{ij} \sum_{m=1}^{N} a_{jm}a_{mi} \tag{1.3}$$

O valor de C_i varia no intervalo 0 e 1 ($0 \geq C_i \geq 1$), onde $C_i = 0$, significa que nenhum dos elementos vizinhos do vértice *i* se interagem entre eles, caso $C_i = 1$ todos os elementos vizinhos interagem entre si. Assim como na conectividade, é possível calcular o coeficiente médio de clusterização da rede, dado pela Equação 1.4:

$$\overline{C} = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{1.4}$$

1.5.1.3 *Betweenness*

Betweenness é a representação da quantidade de caminhos mais curtos que passam por um determinado vértice, sendo ele determinado por:

$$Bet(n) = \sum_{i \neq n \neq j} \frac{\sigma_{ij}(n)}{\sigma_{ij}}, \tag{1.5}$$

onde *n* é o vértice a ser cálculado, σ_{ij} representa todos os possíveis caminhos entre os nós *i* e *j*, e $\sigma_{ij}(n)$ é o total de caminhos mais curtos que passam por *n*.

Um vértice que apresenta um alto valor de *betweenness* são considerados *bottleneck* ou gargalos, isto significa que uma grande quantidade de rotas/caminhos passam por esse vértice.

1.5.1.4 *Closeness*

Closeness representa o caminho mais curto entre um determinado vértice *n* com todos os outros vértices da rede, sendo ele determinado por:

$$Clo(v) = \frac{1}{\sum w \in v^{dist(v'w)}}, \tag{1.6}$$

onde *v* é o vértice a ser calculado, *w* o caminho mais curto encontrado em $v^{dist(v'w)}$.

Um vértice que apresenta um alto valor de *closeness* apresenta uma tendencia de haver muitos outros vértices próximos a ele.

1.6 Bancos de dados

1.6.1 STRING

STRING (*Search Tool for the Retrieval Interactiong Gene/Proteins*) é um banco de dados dedicado para interações funcionais de proteínas em escala global, mantido pelo Laboratório Europeu de Biologia Molecular (*European Laboratory for Molecular Biology* - EMBL) desde 2000 e possui como colaboradores *Swiss Institute of Bioinformatics, University of Zurich, Novo Nordisk Foundation Center for Protein Research* e *Technical University Dresden*,(SZKLARCZYK et al., 2015). Encontra-se na versão 10.0. Está disponível no *site*: `http://string-db.org/`. Atualmente o STRING contém informações de 2031 organismos, mais de 9,5 milhões de proteínas catalogadas e mais de 930 milhões de interações proteicas. Estes dados foram previamente avaliados utilizando como base o repertório do *Kyoto Encyclopedia of Genes and Genomes* (KEGG), onde ele fornece um mapa de rotas metabólicas curadas/verificadas manualmente. O banco de dados contém informações de diferentes métodos de predição de interação/associação: *Conserved Neighborhood, Co-ocurrence, Fusion, Co-expression, Experiments, Databases* e *Textmining*.

- ***Conserved Neighborhood***: Identifica fragmentos de genoma (procariotos) distintos, que possuem a codificação de genes vizinhos semelhantes.
- ***Co-ocurrence***: Verifica a presença e ausência de proteínas ligadas entre diferentes organismos.
- ***Fusion***: Analisa eventos de fusão de genes em cada organismo.
- ***Co-expression***: Evidencia pares de genes coexpressos em um mesmo ou em outros organismos.
- ***Experiments***: Lista de interação proteica retiradas de outros bancos de dados.
- ***Databases***: Lista de interações proteicas extraídas de outros bancos de dados de rotas metabólicas, onde seus dados foram manualmente curados.
- ***Textmining***: Lista de interações proteicas retiradas de resumos dos artigos científicos através de mineração de dados utilizando dicionário específico.

O STRING é capaz de combinar os métodos para a geração de uma rede de interação proteica, pela determinação de um *score*, através do cálculo

$$S = 1 - \prod_{i=1}^{n}(1 - S_i), \tag{1.7}$$

onde S é o *score* ponderado sobre os métodos escolhidos de predição n, S_i é o *score* de cada método escolhido.

1.7 *Gene Ontology*

O projeto *Gene Ontology*(GO), disponível no *site*: `http://geneontology.org/` tem como objetivo o desenvolvimento de identificadores únicos para classificar e organizar genes e seus respectivos produtos, além de centralizar as informações dos mesmos. O GO iniciou seu projeto em 1998 com apenas 3 colaboradores/organismos: *FlyBase* (*Drosophila melanogaster*), *Saccharomyces Genome Database* (*Saccharomyces cerevisiae*) e *Mouse Genome Informatics*(*mus musculus*). Desde então a quantidade de organismos vem aumentando e atualmente o projeto conta com 104 organismos distintos (CONSORTIUM, 2015).

As informações de cada organismo foram organizadas em três classificações independentes, chamadas ontologias, que utilizam critérios distintos. Cada gene e seu respectivo produto está presente nas três ontologias.

1.7.1 Ontologias

- **Componente Celular** (CC): Descrevem as localizações em que os produtos dos genes atuam, sendo elas estruturas subcelulares (organelas), como ribossomos, vacúolos, mitocôndrias, etc, até em complexos macromoleculares, como complexo piruvato desidrogenase, complexo de reparo de DNA, complexo de cadeia respiratória, etc.

- **Função Molecular** (MF): Descrevem atividades ao nível molecular, tais como atividades catalíticas ou atividades de ligação. As MF representam as atividades realizadas pelos genes, entretanto não especificam onde, quando ou em que contexto. Exemplo: ligação ao receptor de insulina, ligação à ativina, etc.

- **Processo Biológico** (BP): Descrevem metas biológicas realizadas por uma ou mais funções moleculares, como por exemplo, reparo de DNA, processo metabólico de ácido úrico, etc.

Por padrão, cada ontologia apresenta ontologias mães e filhas (Figura 1.10), onde as ontoligias mães são as mais abrangentes e as filhas as ontologias mais especifícas, sendo que uma ontologia pode apresentar uma ou mais ontologias mães ou filhas, e também apresentar ontologias mães e filhas em comum com outras ontologias. Com isso, as ontologias apresentaram um grafo direcionado não-cíclico, ou seja, um grafo em árvore.

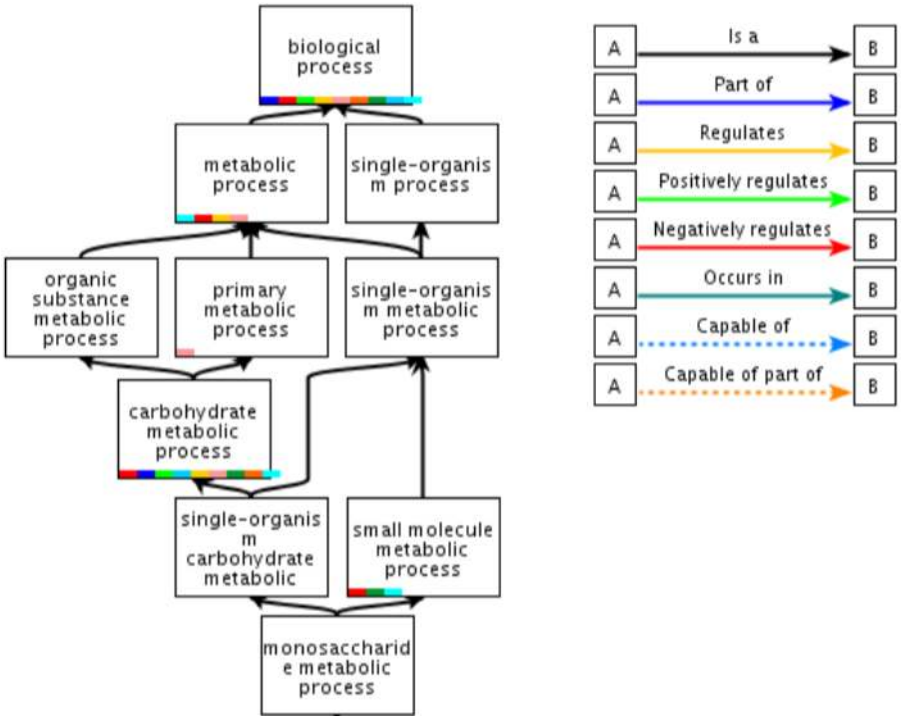

Figura 1.10: Exemplo de ontologia em forma de grafo, onde o *metabolic process* têm como ontologias filhas o *organic substance metabolic process*, *primary metabolic process* e *single-organism metabolic process*. Já a ontologia *monosaccharide metabolic process* têm como ontologias mães o *single-organism carbohydrate metabolic* e *small molecule metabolic process*. Retirado de `https://www.ebi.ac.uk/QuickGO/`

1.8 *Gene Expression Omnibus*

O Gene Expression Omnibus (GEO) (BARRETT et al., 2013) é o maior repositório público de dados de expressão gênica obtidos através de experimentos de alto rendimento, como microarranjo e sequenciadores de nova geração, que a comunidade científica disponibiliza. É mantido pelo *National Center for Biotechnology Information*(NCBI),disponível no *site*: `http://www.ncbi.nlm.nih.gov/geo/`. Nesse repositório está contido uma variedade de tipos de dados, como metadados, dados processados e até dados brutos, fazendo com que atualmente tenha dados de mais de 3,6 mil organismos, mais de 72 mil experimentos, que contemplam mais de 1,8 milhões de amostras depositadas de mais de 16 mil plataformas distintas.

2 *Objetivos*

O presente trabalho visa a implementação de uma nova metodologia de integração de dados de interação proteína-proteína e níveis de expressão gênica para uma melhor mensuração global do metabolismo de uma célula, além de auxiliar na busca de assinaturas transcricionais para tumores.

2.1 Objetivos Específicos

- Implementação de algoritmos de clusterização de redes;
- Automação de processos de análise de enriquecimento funcional e expressão gênica;
- Automação da compilação das informações obtidas em formato de gráfico;
- Aplicação da metodologia em amostras de câncer;
- Comparação com metodologias de análise de enriquecimento funcional.

3 *Material e Métodos*

3.1 *Workflow*

A figura 3.1 apresenta todas as etapas para a obtenção da integração de redes e transcriptoma (aqui chamado de transcriptograma), para isso foram necessárias 2 linguagens de programação: C e R. A linguagem C foi utilizada para realizar um processamento de dados massiva. A linguagem R foi utilizada para o processamento de uma menor quantidade de dados em comparação com a linguagem C. Para uma melhor compreensão dos passos realizados pelo *workflow*, cada segmento do mesmo está divido em subseções para uma melhor explicação.

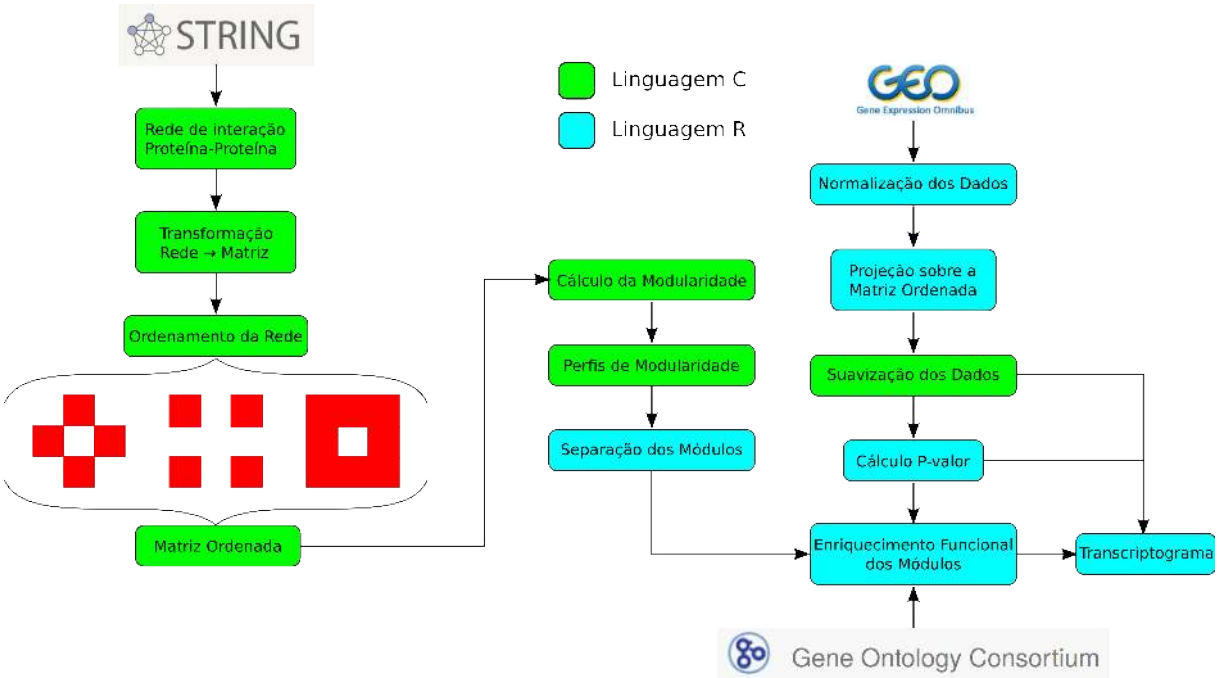

Figura 3.1: *Workflow* referente as etapas para a obtenção do transcriptograma. As caixas em verde representam os algoritmos que utilizam a linguagem de programação C enquanto que as caixas em ciano são os algoritmos que utilizam a linguagem de programação R.

3.2 Ordenamento

A figura 3.2 apresenta a etapa da obtenção da rede e seu respectivo ordenamento. Para isso é necessário a obtenção de uma rede no banco de dados STRING, essa rede possui uma forma bi-dimensional que é transformada na forma de uma matriz adjacente. Logo após a transformação é realizado o ordenamento da rede, onde pode-se escolher um dos três modelos de análise de vizinhança: “Cruz”, “X” ou “Anel”. Ao final temos uma matriz ordenada.

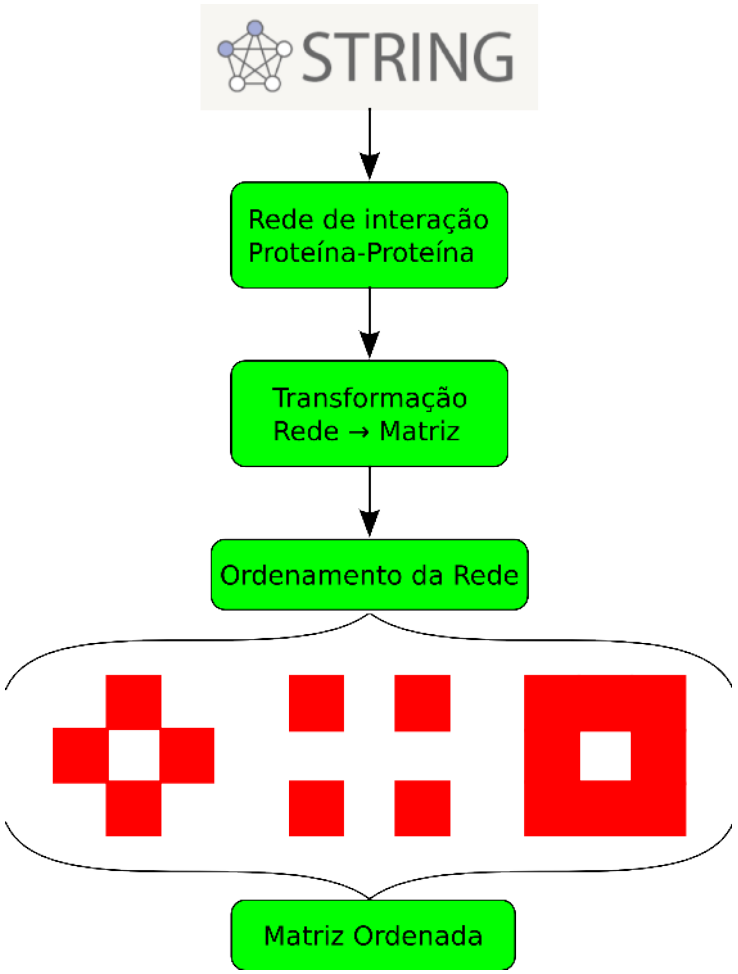

Figura 3.2: Segmento do workflow referente a etapa de ordenamento da rede.

A partir de uma rede de interação, é identificado cada elemento dela como um vértice V , onde para cada vértice é atribuído um valor numérico inteiro, iniciando em 0 até o máximo de elementos, criando então uma legenda. A partir dessa nova nomenclatura cria-se uma matriz de adjacência M de N linhas e N colunas. As interações entre os vértices V são chamadas de arestas A. Para cada interação dos elementos dessa rede é atribuida um valor Verdadeiro (V) no respectivo elemento m_{ij} da matriz de adjacência M, caso contrário será atribuído um valor Falso (F) (Figura 3.3). A atribuição dos valores Verdadeiro e Falso na matriz de adjacência, ou seja

valores booleanos, torna a matriz mais simples e rápida de ser interpretada pelo processador do computador, aumentando assim a sua performance, além de diminuir o custo de memória RAM necessária para a construção da matriz M, pois um valor booleano consome 8 bits de memória, enquanto um valor inteiro consome no mínimo 16 bits, ou seja, diminuí-se pela metade o custo de memória.

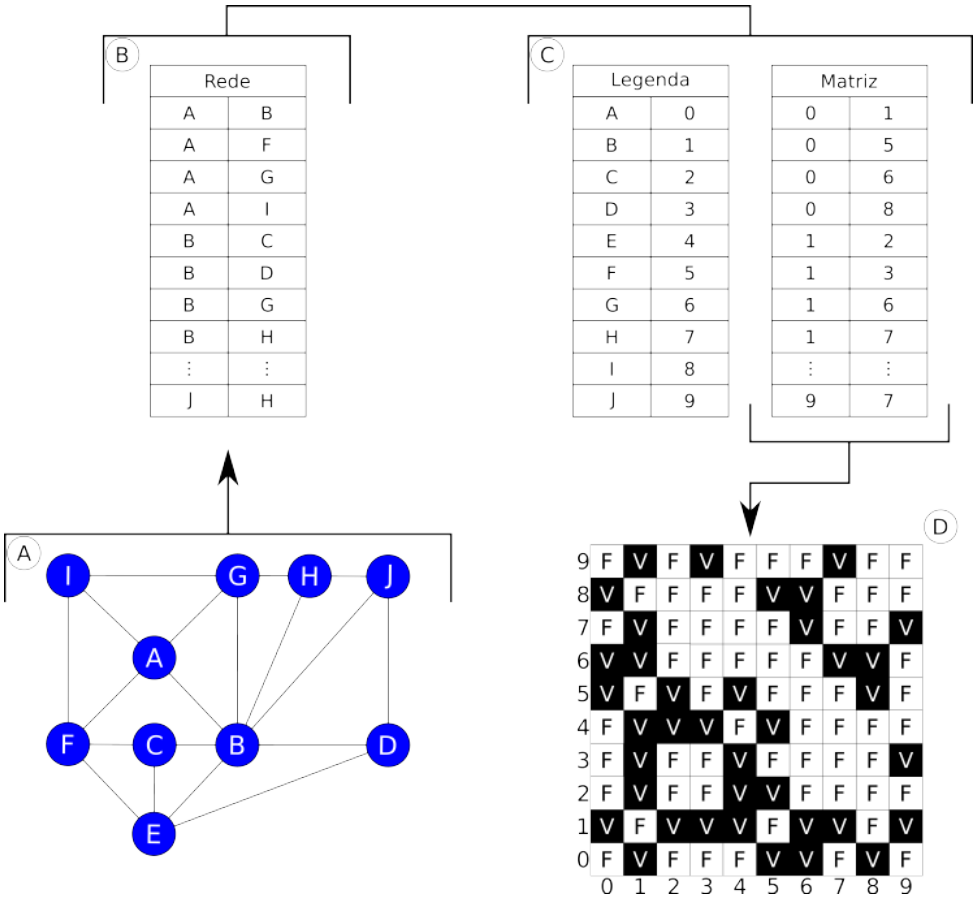

Figura 3.3: Transformação da rede de interação em uma matriz de adjacência booleana. (A)Exemplo de uma rede biológica, onde os vértices representam as proteínas, as letras representam os nomes dos elementos e as arestas representam a interação entre eles; (B) Representação da rede de interação em forma de lista, onde é apresentado somente os elementos que possuem arestas (interagem); (C) Enumeração dos elementos da lista e criação da legenda, para que não haja a perda dos nomes dos elementos; (D) Transformação da lista enumerada em matriz adjacente, em que a existência da interação entre os elementos será inserido um valor Verdadeiro (V), caso contrario será inserido um valor Falso (F).

A identificação dos vértices, por padrão, seguirá de forma crescente para cada novo elemento identificado. Entretanto essa ordem pode ser alterada através da permutação das linhas e colunas, e a cada permutação uma nova matriz de adjacência é criada, sendo assim, o número de possíveis combinações segue a ordem de $N!$ (fatorial de N). Sabendo dessa possibilidade, é analisado novos ordenamentos para obter uma matriz que possua a menor quantidade possível

de elementos Falsos próximos a diagonal principal, fazendo com que seja identificados grupos interagentes. Por meio desta técnica podemos identificar os grupos de elementos que possuam uma grande interação entre si e além de estarem próximos de grupos distintos que tenham grande interação entre si. Porém, para se obter essas características têm-se que tentar satisfazer duas condições:

- **Condição 1:** Quanto menor a quantidade de elementos não-interagentes e interagentes dispersos melhor a configuração, como apresentado na Figura 3.4. Existem diversas configurações possíveis para cada elemento central, mas a pior configuração é aquela em que o elemento central não possuí nenhum elemento vizinho, e conforme exista mais elementos vizinhos melhor a configuração até chegar no máximo que são 8 elementos vizinhos entorno do elemento central.

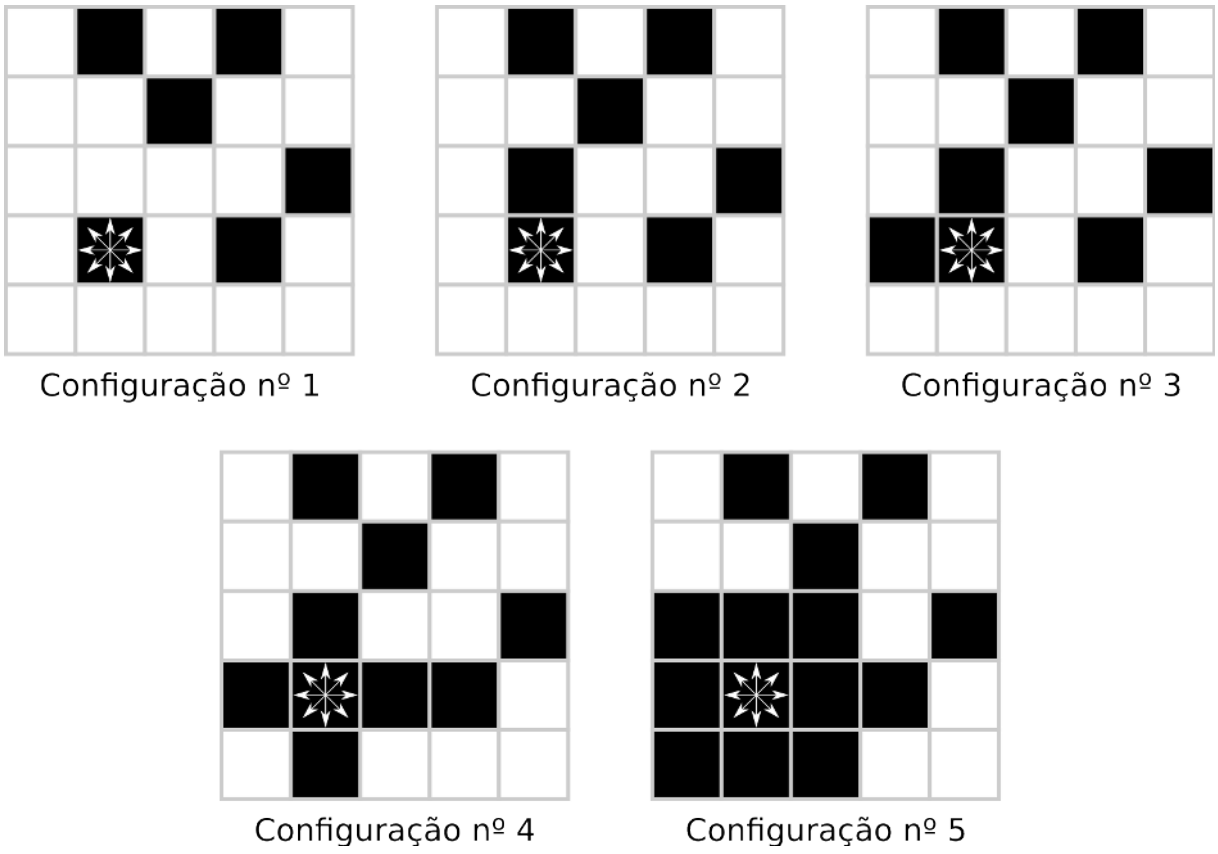

Figura 3.4: Exemplo de cinco possíveis configurações de vizinhanças em relação ao elemento central da matriz de adjacência. O elemento central é representado pelo quadrado preto com setas brancas: a configuração nº 1 é a configuração menos favorável, pois não há nenhum elemento interagente vizinho do elemento central ; já a configuração nº 5 é a onfiguração mais favorável, pois todos os vizinhos do elemento central são interagentes.

- **Condição 2:** Quanto mais próximo o elemento central está da diagonal principal melhor a configuração, como demonstrado na Figura 3.5.

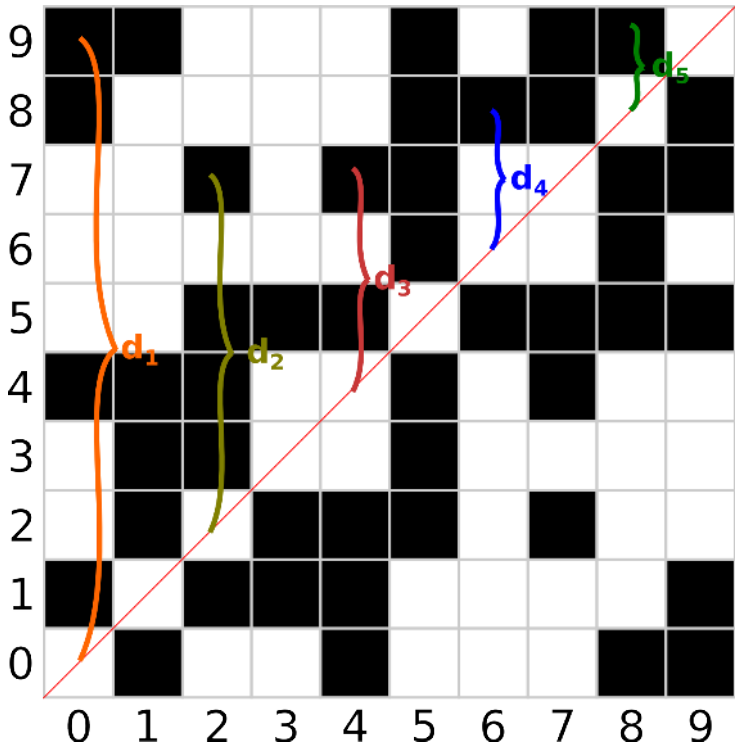

Figura 3.5: Exemplo de cinco possíveis distâncias do elemento central em relação a diagonal principal da matriz de adjacência. A diagonal principal é representada pela linha vermelha na matriz: a distância d_1 é a distância menos favorável, pois o elemento central está mais afastada da diagonal; enquanto que a distância d_5 é a melhor possível, pois o elemento central está mais próximo possível da diagonal.

Assim como a primeira condição, a segunda possui diversas configurações, mas a menos favorável é quando o elemento central está o mais distante possível da diagonal principal. A melhor configuração é quando o elemento central se aproxima da diagonal principal até possuir distância de valor igual a 1. Entretanto essas condições não são concordantes a todo momento, pois ao fazer a permutação de vértices, é feito com que os elementos Verdadeiros fiquem mais próximos a diagonal porém, ao mesmo tempo, pode-se criar uma dispersão dos outros elementos.

3.2.1 Modelo Cruz

Para que as duas condições citadas anteriormente sejam atingidas, utiliza-se o método proposto por (RYBARCZYK-FILHO et al., 2011), que para atender a condição de menor quantidade de elementos dispersos, soma-se ao elemento central a quantidade de vizinhos Falsos

próximos, onde os vizinhos analisados são os acima, abaixo, à direita e à esquerda do elemento central. Esse tipo de análise de vizinhos foi nomeado como modelo "cruz" (Figura 3.6). E para atender a condição de proximidade à diagonal, multiplica-se o elemento central pela sua distância ($d_{i,j}$) em relação a diagonal principal.

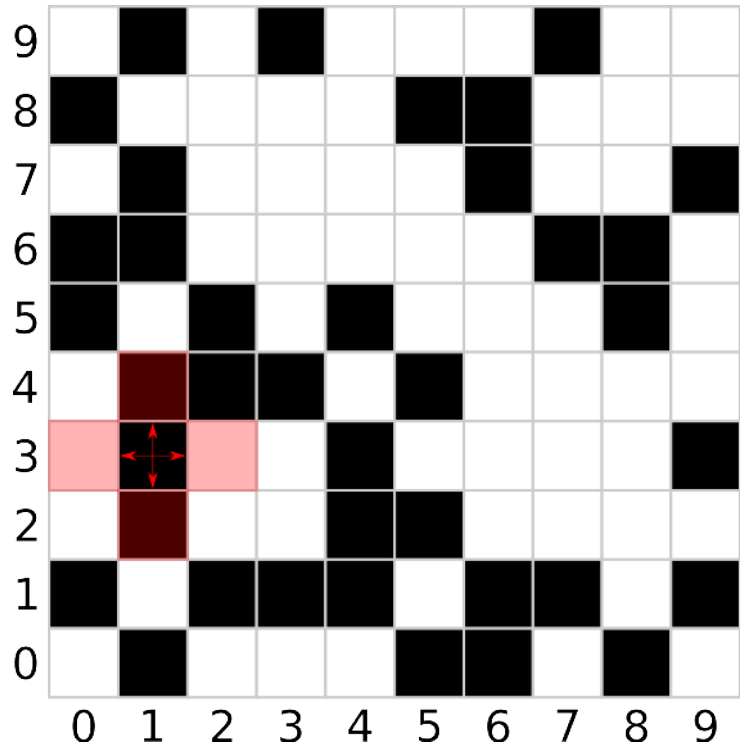

Figura 3.6: Análise de vizinhança do modelo "cruz", onde é selecionado um elemento central Verdadeiro (V), representado pelo quadrado preto com as flechas vermelhas, e por meio dele são analisados os elementos de cima, baixo, esquerda e direta, representados pelos quadrados vermelhos.

Realiza-se esse cálculo em todos os elementos Verdadeiros da matriz e soma-se seus resultados. Com isso obtém-se o custo energético (ε) da matriz na configuração atual, mas para reduzir esse custo energético é necessário a aplicar o Método de Monte Carlo para minimizar ε através do reordenamento da matriz e do modelo "cruz" para analisar a vizinhança (Equação 3.1).

$$\varepsilon = \sum_{j=1}^{V}\sum_{i=1}^{V} d_{i,j}\{|m_{i,j} - m_{i+1,j}| + |m_{i,j} - m_{i-1,j}| + |m_{i,j} - m_{i,j+1}| + |m_{i,j} - m_{i,j-1}|\} \qquad (3.1)$$

Para que o método seja eficiente, o primeiro passo é a aleatorização dos vértices da matriz M. Ao permutar uma coluna, a linha correspondente deve-se obrigatoriamente ser permutada na mesma posição. Isso é feito para que nenhuma interação seja perdida ou criada, com isso aplica-se a função anterior e obtém-se ε_i. O próximo passo é permutar aleatoriamente dois

vértices, obtendo assim uma nova configuração (Figura 3.7).

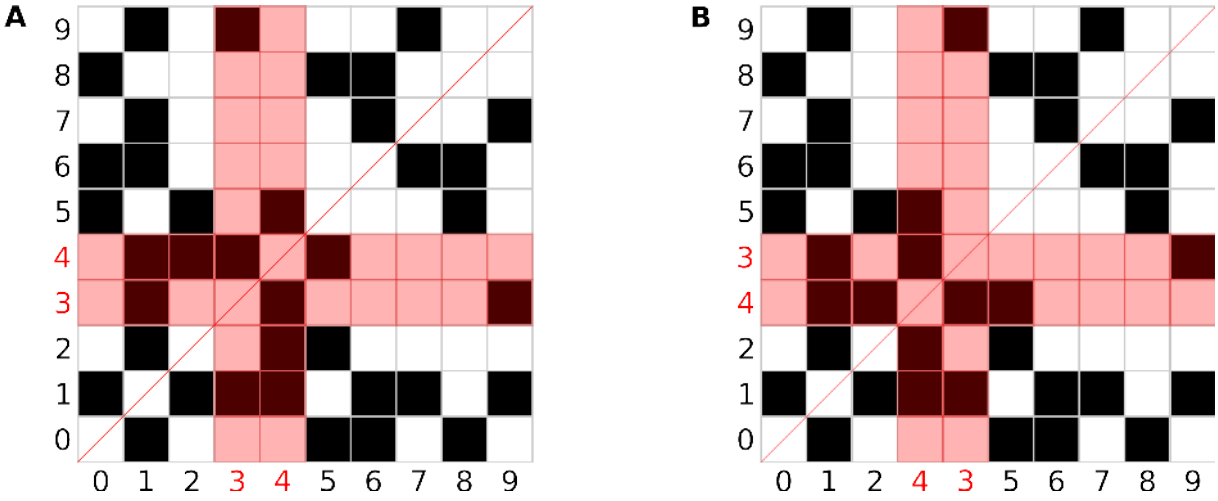

Figura 3.7: Permutação de vértices da matriz adjacente para criação de uma nova configuração da matriz. **(A)** Matriz adjacente em uma configuração aceita pelo sistema, onde os vértices, as linhas e colunas em vermelho são os escolhidos para realizar a permutação, linha com linha e coluna com coluna. **(B)** Matriz adjacente após a permutação dos vértices da matriz, representados pelos vértices, linhas e colunas em vermelho.

Aplica-se novamente a função e obtêm-se ε_f. Obtendo ambas as energias (ε_i e ε_f), é verificado se $\varepsilon_f < \varepsilon_i$, caso seja, a nova configuração da matriz é aceita e a permutação dos vértices é realizada novamente. Caso contrário, $\varepsilon_f > \varepsilon_i$, a nova configuração é aceita com uma probabilidade de $\exp[-(\varepsilon_f - \varepsilon_i)/T]$, sendo T um parâmetro que simula a temperatura no Método de Monte Carlo.

Para a simulação da temperatura T, inicia-se com um valor T_0 muito elevado, sempre em relação ao primeiro custo energético ε obtido após a aleatorização, e ao longo das análises diminuí-se esse valor. Para cada decréscimo de T, percorre-se uma determinada quantidade de passos de Monte Carlo necessário para atingir o equilíbrio termodinâmico, e então é reajustado o seu valor através de uma função de arrefecimento $T^* = \mu T$, sendo $0 < \mu < 1$ e T^* a nova temperatura. Ao utilizar essa técnica, conhecida como *simulated annealing*, é possível ultrapassar os mínimos locais atingindo valores mais elevados a fim de atingir valores mais baixos de energia até um encontrar um mínimo absoluto (Figura 3.8).

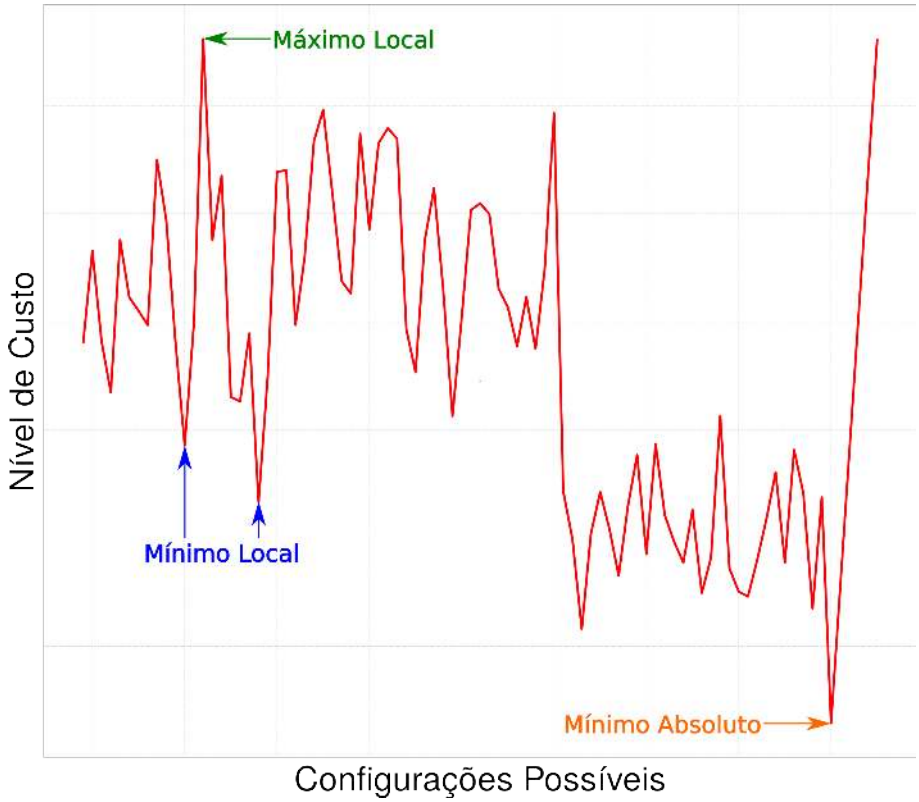

Figura 3.8: Perfil energético em função de todas as configurações possíveis de ordenamento.

3.2.2 Modelo X

A partir do modelo "Cruz" desenvolveu-se dois novos modelos, o modelo "X" , utiliza como base a condição de menor quantidade de elementos dispersos (MOLAN; RYBARCZYK-FILHO, 2014). Entretanto o que altera em relação ao modelo anterior é a vizinhança analisada. Nesse modelo os vizinhos analisados são as diagonais superior direita e esquerda, e as diagonais inferior direita e esquerda (Figura 3.9).Por consequência dessa alteração tem-se uma nova função de custo energético (Equação 3.2).

$$\varepsilon = \sum_{j=1}^{A}\sum_{i=1}^{A} d_{ij}\{|m_{i,j} - m_{i-1,j+1}| + |m_{i,j} - m_{i-1,j-1}| + |m_{i,j} - m_{i+1,j+1}| + |m_{i,j} - m_{i+1,j-1}|\} \tag{3.2}$$

3.2.3 Modelo Anel

O segundo modelo criado por Molan *et al* (MOLAN; RYBARCZYK-FILHO, 2014) é o "Anel", que assim como o modelo "X" utiliza como base a condição de menor quantidade

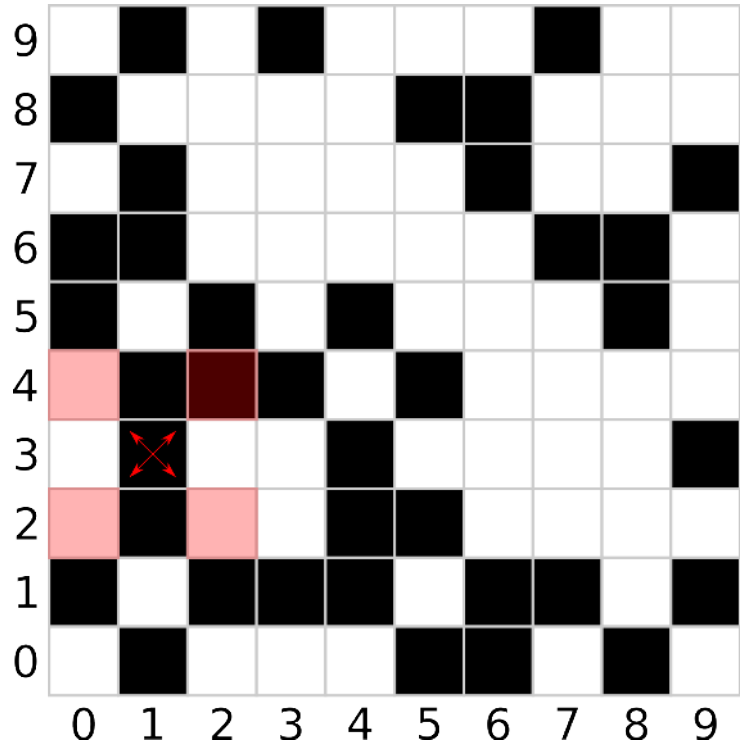

Figura 3.9: Análise de vizinhança do modelo “X”, onde ? selecionado um elemento central Verdadeiro (V), representado pelo quadrado preto com as flechas vermelhas, e por meio dele são analisados os elementos superiores esquerda e direita e inferiores esquerda e direita, representados pelos quadrados vermelhos.

de elementos dispersos, porém entre os três modelos ele é o mais completo abrangendo uma maior quantidade de vizinhos, ou seja, analisa todos os oito vizinhos do elemento central (Figura 3.10).E assim como no modelo ”X´´, é necessário uma nova função de custo energético (Equação 3.3).

$$\varepsilon = \sum_{j=1}^{A}\sum_{i=1}^{A} d_{i,j}\{|m_{i,j} - m_{i-1,j-1}| + |m_{i,j} - m_{i-1,j}| + |m_{i,j} - m_{i-1,j+1}| + |m_{i,j} - m_{i,j+1}| \\ + |m_{i,j} - m_{i+1,j+1}| + |m_{i,j} - mi + 1, j| + |mi, j - m_{i+1,j-1}| + |mi, j - mi, j - 1|\} \tag{3.3}$$

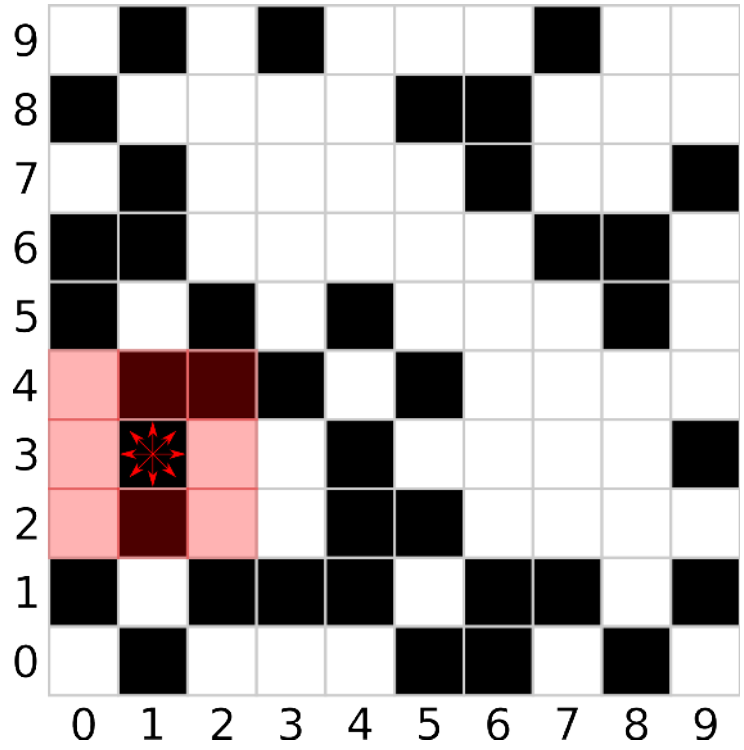

Figura 3.10: Análise de vizinhança do modelo “Anel”, onde é selecionado um elemento central Verdadeiro, representado pelo quadrado preto com as flechas vermelhas, e por meio dele são analisados os elementos de cima, baixo, esquerda, direita, superiores esquerda e direita e inferiores esquerda e direita, representados pelos quadrados vermelhos.

3.2.4 Alterações no Método de Clusterização

O primeiro método foi desenvolvido por Rybarczyk-Filho *et al* (RYBARCZYK-FILHO et al., 2011) utilizando somente o modelo “cruz”, ou seja, criava-se uma matriz adjacente e permutava todas as interações dos elementos, fazendo com que o processo fosse demorado. Outra informação é que foi utilizada a linguagem de programação Fortran.

O segundo método foi o desenvolvido por Molan *et al* (MOLAN; RYBARCZYK-FILHO, 2014), que no lugar de usar uma matriz adjacente, utilizava somente a uma lista com as interações, que é basicamente um passo anterior ao da criação da matriz adjacente, por meio desta técnica foi capaz de reduzir o tempo do processo e também reduzir o numero de trocas.Além disso, foram desenvolvidos os dois outros modelos de ordenamento o “X” e “cruz”, e a linguagem de programação utilizada foi C++.

O terceiro método foi o desenvolvido por (KUENTZER, 2014), que continuou utilizando a matriz adjacente, porém adicionou novas técnicas junto a matriz adjacente, tais como: um vetor de ponteiros para troca, a utilização parcial dos vértices, utilização de somente os vértices da diagonal principal e a criação de uma nova representação da matriz. O vetor de ponteiros foi

utilizado para armazenar a nova posição do vértice (Figura 3.11 C). A utilização parcial dos vértices (Figura 3.11 A) associado somente os vértices da parte superior da diagonal principal (Figura 3.11 B) reduz a quantidade necessária de cálculo, pois por meio dessas modificações é cálculado apenas os vértices que, possivelmente, sofreram alguma alteração. E por último a adição de uma nova representação da matriz (Figura 3.11 D), além da matriz adjacente, que basicamente é uma matriz que apresenta somente a interação entre os vértices, ao utilizar esta técnica é mais fácil a localização da interação, porém é mais complicado para encontrar vizinhos associados.

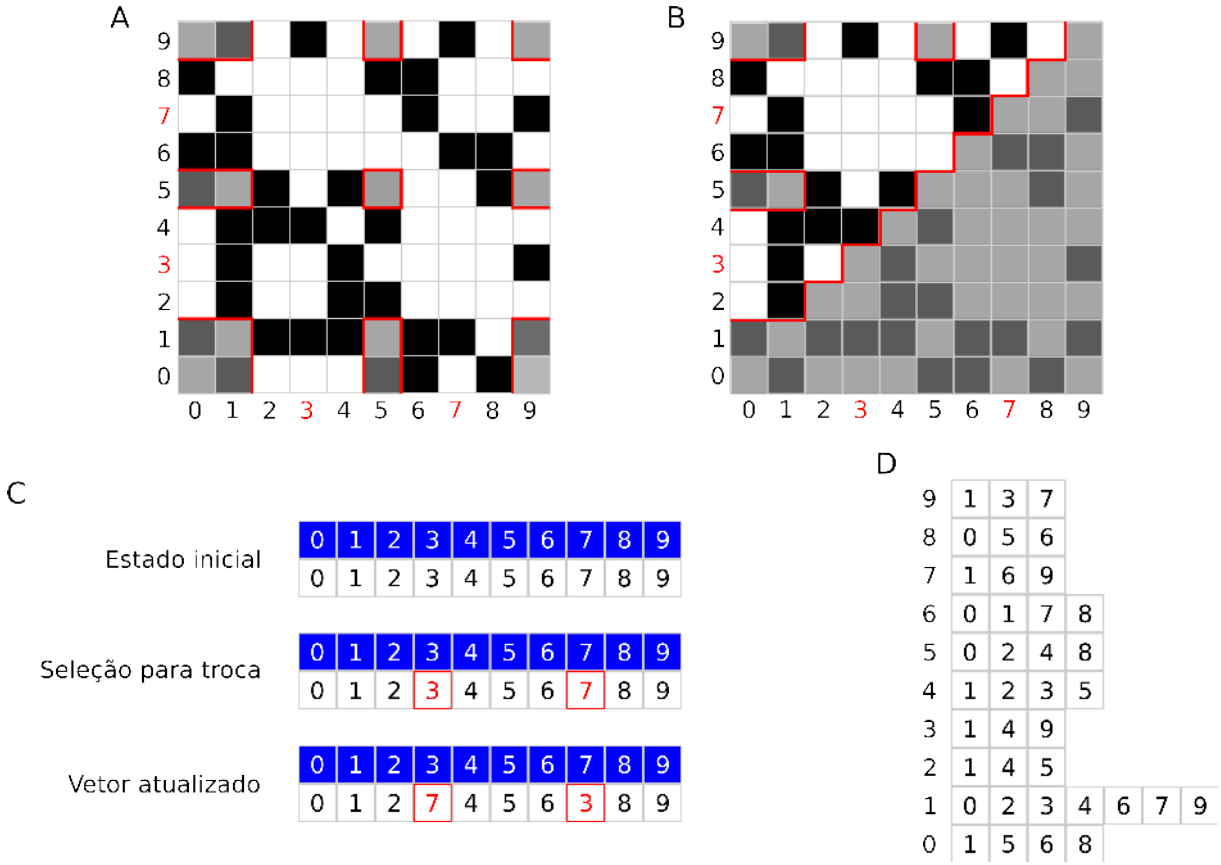

Figura 3.11: Representação gráfica das alterações feitas por Kuentzer *el al*. **(A)**. Foram selecionados os índices 3 e 7 para fazer a permutação, com isso somente os índices da esquerda (2 e 6) e direita (4 e 8) além dos índices permutados serão avaliados. Sendo todo o restante, em cinza, não sendo necessário o seu cálculo por não haver qualquer tipo de alteração. **(B)**. Complementação da figura anterior, onde toda a parte inferior da diagonal principal da matriz não precisa ser analisada, pois todas as interações necessárias estão presentes na parte superior da diagonal, isso só é possível pois a matriz é simétrica. **(C)**. Representação de como é realizado as trocas através do vetor de ponteiros, onde a parte azul representa os locais de cada índice, e a parte branca a posição atual dos índices da matriz. Ao selecionar o índice 3 e 7 permutamos seus lugares fazendo com que o índice 7 agora verifique a vizinhança nos índices 2 e 4, assim como o índice 3 verifica a vizinhança nos índices 6 e 8. **(D)**. Representação da matriz possuíndo apenas as interações, como por exemplo o índice 0 interage com os índices 1, 5, 6 e 8 como é possível verificar na matriz anterior.

Com essas alterações o tempo gasto para o reduziu drasticamente, entretanto o modelo utilizado por Kuentzer *et al* (KUENTZER, 2014) foi somente o "cruz". A linguagem de programação utilizada foi umas das variações da linguagem C. O método desenvolvido nesse trabalho tomou como base o método do (KUENTZER, 2014), porém foram adicionados duas novas técnicas: dicionário e a matriz adjacente do tipo booleano. A utilização do dicionário trouxe uma maior facilidade para a busca das informações armazenadas, e com a utilização do valor booleano, além da diminuição do consumo de memória RAM, o tempo de processamento pode ser reduzido por apresentar uma forma mais próxima a da linguagem de máquina. Além disso, foram adicionados os dois modelos desenvolvidos por (MOLAN; RYBARCZYK-FILHO, 2014) e um sistema de verificação do andamento do ordenamento. A linguagem de programação utilizada foi a linguagem C.

3.3 Modularidade

A figura 3.12 representa a etapa referente a obtenção de perfis de modularidade e separação de módulos. Para isso é necessário que a matriz esteja reordenada, através dessa matriz será aplicada um cálculo para a extração da modularidade da rede.

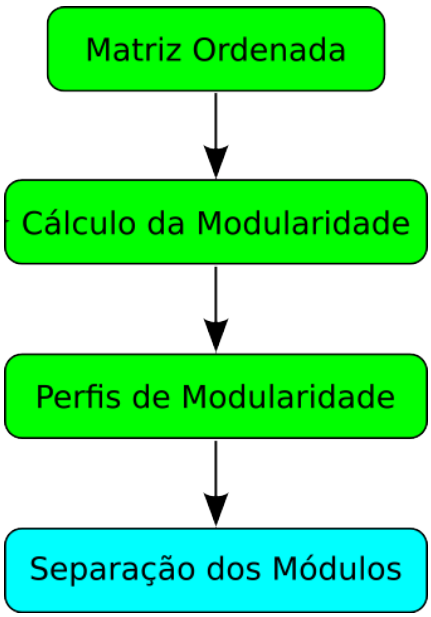

Figura 3.12: *Workflow* referente a etapa de modularidade da rede.

Modularidade é uma técnica capaz de auxiliar na verificação de clusters de vértices para posteriormente obter as ontologias presentes em uma rede, por meio da análise de interação entre os elementos da rede em um determinado intervalo. A técnica busca módulos de interação

obtidos pelo reordenamento da matriz. Somente esses módulos não são capazes de identificar se há ontologias associadas, entretanto é um passo fundamental para a sua identificação.

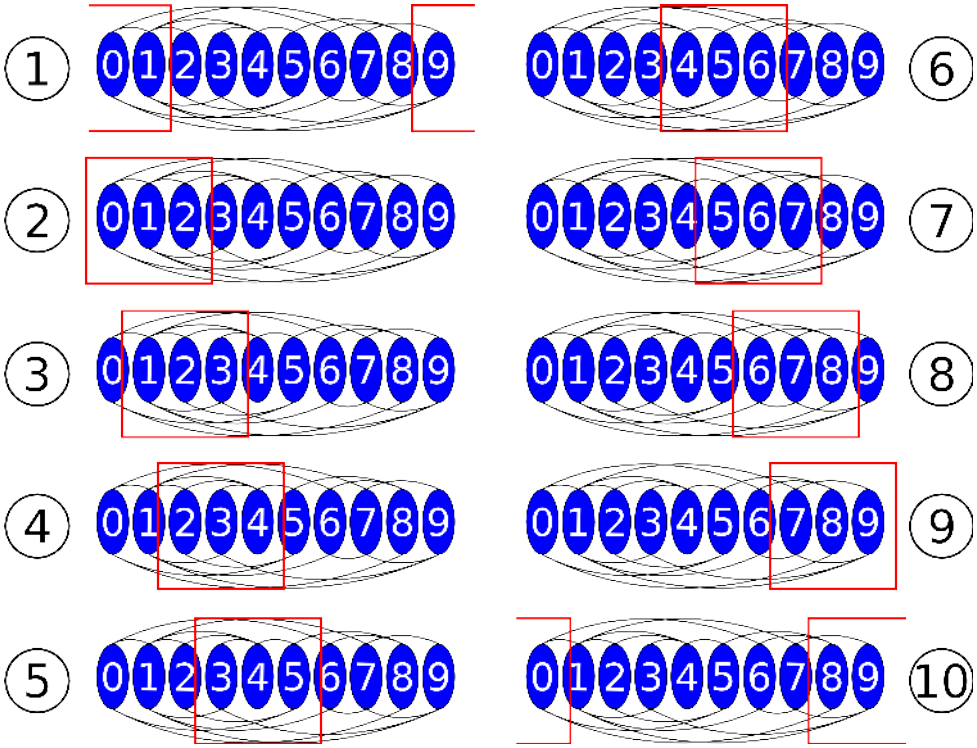

Figura 3.13: Cálculo de Modularidade para obtenção de módulos de interação da rede. As elipses representam os elementos da rede, os arcos pretos são as interações entre os elementos e o quadrado vermelho é o intervalo (janela) de elementos a serem analisados.

Para identificar os módulos de interação, utiliza-se uma função que verifica a quantidade de interações entre os elementos contidos em uma janela em razão da quantidade total de interações na rede. A janela é o intervalo de elementos para a análise dos módulos, a mesma tem um tamanho mínimo de 3 e um tamanho máximo de elementos da rede, essa janela é sempre de valor ímpar, pois o resultado da função sempre será inserido no elemento central da janela. A janela definida deve passar por todos os elementos da rede até que todos os elementos dessa rede possua um nível de modularidade,e a partir disso obtém-se o perfil de modularidade (Figura 3.13). Por meio deles é possível a verificação da interatividade dos módulos, onde módulos muito interativos são obtidos através de janelas de valores próximos a quantidade máxima de elementos na rede, apesar disso eles não são capazes de demonstrar resultados de qualidade, e conforme diminui o valor da janela surgem inúmeros módulos sem informação relevante.

3.3.1 Separação dos Módulos

Após a obtenção da modularidade utiliza-se uma ferramenta chamada shiny da linguagem R, que uma das suas funções é a possibilidade de trabalhar com gráficos de forma interativa. Através dela é possível a separação dos módulos de forma manual (Figura 3.14).

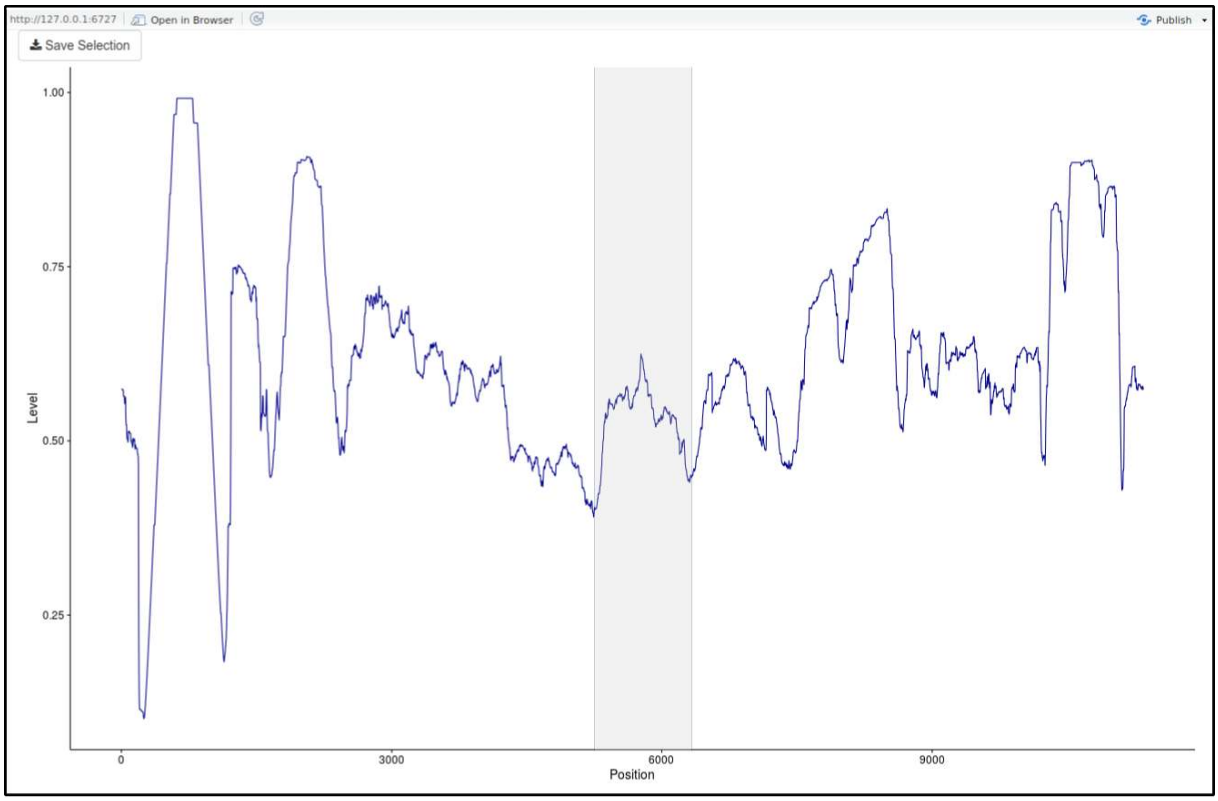

Figura 3.14: Interface gráfica construída com o uso do shiny para separação dos módulos, onde a linha azul é a posição relativa da proteína em relação ao nível de modularidade, a caixa cinza são as proteínas do módulo selecionado pelo usuário.

3.4 Análise de Expressão Gênica

A figura 3.15 apresenta a etapa referente a análise dos dados de expressão gênica. Para isso é necessário a obtenção de uma série no banco de dados GEO, que por sua vez é aplicado uma técnica de normalização para a extração dos dados dos *chips*. Logo após é projetado sobre a matriz ordenada para análisar apenas os elementos presentes na rede e por fim é realizada a suavização dos dados, ou seja, uma média e o cálculo do p-valor referente as amostras.

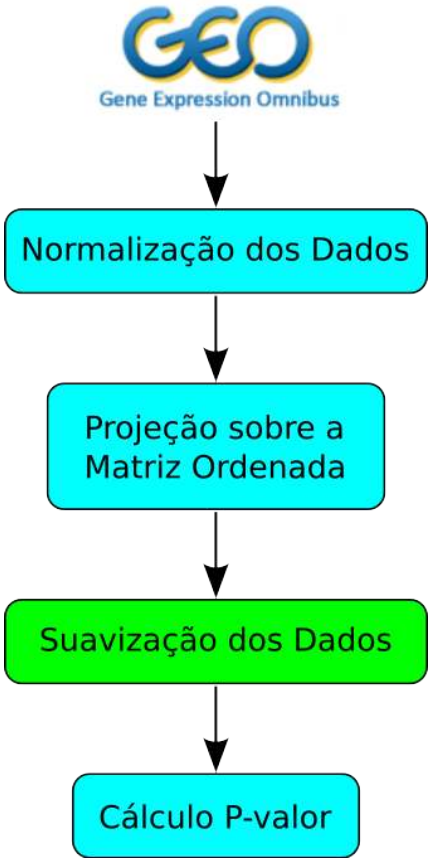

Figura 3.15: *workflow* referente a etapa de análise de expressão gênica

3.4.1 Normalização dos Dados

Para a normalização dos dados de expressão obtidos no GEO, é necessário a utilização de uma técnica de normalização, foi optado pela técnica RMA (*Robust Mult-array Analysis*), pois a técnica apresenta uma melhor precisão e sensibilidade em comparação as demais técnicas MAS5 e GCRMA (IRIZARRY et al., 2003a). Entretanto um dos requisitos da técnica é a exigência de no minímo triplicatas de cada amostra para que a técnica apresente seu potencial máximo.

3.4.2 Projeção sobre a Matriz Ordenada

Com os dados normalizados, analisa-se quais das sondas estão presentes na rede proteica, caso alguma sonda não possua a proteína presente na rede, a mesma é descartada, e as que possuem mais de uma sonda por proteína é realizada uma média da expressão dessa sondas.

3.4.3 Suavização dos Dados

Aplica-se o cálculo similar ao de modularidade em cada microarranjo, porém a alteração é que no lugar da razão é utilizado a média, pois é realizada a soma dos valores de expressão de cada sonda e divido pelo tamanho da janela definida. A janela utilizada nesse cálculo deve ser a mesma utilizada para a obtenção do perfil de modularidade.

Com a finalização do cálculo dos dados de expressão, passa-se a realizar a suavização desses dados, feito a partir da razão das amostras teste pela referência.

3.4.4 Cálculo do p-valor

Ao realizar todos os procedimentos, é necessário a comprovação se realmente um gene/proteína encontra-se super-expresso ou sub-expresso. Calcula-se o p-valor das sondas presentes na rede proteica. Realiza-se esse procedimento utilizando a função beta incompleta (equação 3.4), que realiza uma probabilidade de ocorrência em um intervalo finito. Na utilização da função são necessários os dados dos módulos de expressão antes da aplicação da suavização, que servem como parâmetros para que a função retorne os valores de p-valor para cada uma das sondas.

$$B_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt \tag{3.4}$$

3.5 Enriquecimento Funcional

A figura 3.16 apresenta a etapa referente a obtenção dos processos biológicos através do enriquecimento funcional dos módulos separados através do perfil de modularidade e dos genes que apresentaram um p-valor abaixo de um determiado valor. Para isso utiliza-se o banco de dados Gene Ontology Consortium. Após a aquisição dos processos biológicos é combinado as informações da suavização dos dados, p-valor e o enriquecimento funcional para a obtenção do transcriptograma. Com a separação dos módulos do perfil de modularidade e a

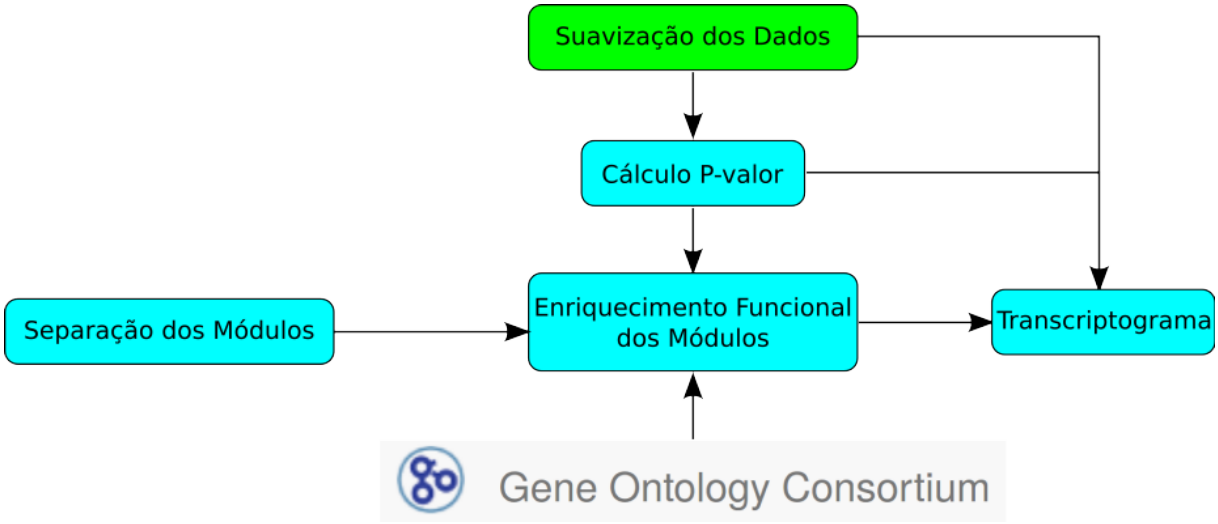

Figura 3.16: *Workflow* referente a etapa de obtenção do enriquecimento funcional.

obtenção do p-valor, realiza-se a separação das proteínas que possuem um p-valor igual o menor a 1×10^{-5}. Com a obtenção das proteínas, projetam-as sobre os módulos e aplicam-se novas ferramentas como biomaRt, GO.db, ClusterProfiler e org.Hs.eg.db (ferramentas da plataforma Bioconductor `http://bioconductor.org/`) para encontrar, em cada módulo, os possíveis processos biológicos alterados. Os processos biológicos considerados alterados são todos aqueles que além de obterem sigfinicância estatística também devem apresentarno mínimo 70% das proteínas do processo biológico contidos nesse módulo.

3.6 *Software* de Análise de Enriquecimento Funcional

3.6.1 *Generally Applicable Gene-set Enrichment* (GAGE)

O GAGE (LUO et al., 2009) é uma metodologia baseada em *Gene set Analysis* (GSEA) que utiliza, em primeira instância, todos os genes disponíveis, não realizando uma pré-filtragem dos genes, além de incorporar rotas metabólicas. Para a realização de seus cálculos, o GAGE necessita de *gene sets* experimentais ou curados/validados. Para a comparação realiza-se a média dos valores absolutos de cada gene, presente no arquivo normalizado, para verificar possíveis genes superexpressos e subexpressos, e então descartar genes que não apresentam um resultado satisfatório, após esta etapa, é determinado o tipo de comparação entre os *chips*, que podem ser um-por-um, um-por-grupo e grupo-por-grupo. Neste trabalho utiliza-se a comparação grupo-por-grupo, no qual foi realizada uma média do nível de expressão dos genes presentes em cada amostra do grupo .

Após a criação de todas as variáveis, citadas anteriormente, a metodologia as utiliza no teste estatístico denomidado teste-t de duas amostras, que apresenta como resposta o resultado do teste, o p-valor e o q-valor de possíveis processos biológicos. Com estes resultados é possível identificar os *gene sets* significativamente alterados. O GAGE pode ser obtido no link: `http://bioconductor.org/packages/gage/`

3.6.2 *Parametric Analysis of Gene set Enrichment* (PAGE)

O PAGE (KIM; VOLSKY, 2005) baseia-se em GSEA, entretanto,as duas metodologias diferem-se na forma de preparo dos dados e o teste estatístico utilizado para a determinação de *gene sets* significativamente alterados.

Para o preparo é realizada uma normalização nos dados do microarranjo, onde a média de expressão de cada chip é tomada com valor igual a 1.000, após essa normalização é verificado se existe algum valor de expressão abaixo de 100, caso exista esse valor é elevado ao expoente 100, e por fim os valores são transformados aplicando um logarítmo de base 2 ($\log_2$).

O teste estatístico utilizado é o teste-z de uma amostra, para que seja aplicado esse teste o PAGE faz a comparação grupo-por-grupo, não sendo possível a alteração do tipo de comparação. O resultado final é o valor do teste estatístico z, o p-valor e o q-valor, assim como o GAGE. O PAGE pode ser obtido no link: `https://github.com/zhilongjia/PAGE`

4 *Resultados e Discussão*

4.1 Comparação entre os modelos para ordenamentos

Foram prospectadas duas redes proteicas do banco de dados STRING com *score* 0,7 e 0,8. Para a constituição desse *score* foram utilizados os métodos *Experiment* e *Database*. A rede com *score* 0,7 possuí 463.208 interações e 11.350 proteínas e a rede com *score* 0,8 possuí 440.400 interações e 10.483 proteínas. A partir dessas redes foram aplicados os três modelos de análise de vizinhança "Cruz", "X" e "Anel", em cada modelo foi utilizado 3.000, 5.000 e 10.000 passos de Monte Carlo. Para fins estatísticos, cada combinação de método e passo de Monte Carlo foi executado 10 vezes. O resultado é apresentado na Tabela 4.1 e 4.2.

Tabela 4.1: Comparação entre as combinações dos modelos e passos de Monte Carlo em relação ao tempo de médio de processamento da rede de *score* 0,7.

Modelo	Passo de Monte Carlo	Tempo Médio (s)	Desvio Padrão (s)
Cruz	3.000	969,2	28,31
	5.000	1636,6	31,09
	10.000	3104,5	56,94
X	3.000	1075,5	24,97
	5.000	1819,3	40,34
	10.000	3448,9	34,74
Anel	3.000	1433,3	39,31
	5.000	2251,8	45,74
	10.000	4294,4	70,91

Tabela 4.2: Comparação entre as combinações dos modelos e passos de Monte Carlo em relação ao tempo de médio de processamento da rede de *score* 0,8.

Modelo	Passo de Monte Carlo	Tempo Médio (s)	Desvio Padrão (s)
Cruz	3.000	960,2	78,04
	5.000	1.382,5	18,57
	10.000	2.683	34,15
X	3.000	979,7	24,86
	5.000	1.532,8	10,22
	10.000	2.965	31,48
Anel	3.000	1.244,3	45,54
	5.000	1.895,5	16,65
	10.000	3.683,4	61,58

Ao observar os resultados, podemos verificar que o modelo “Cruz” apresentou menor tempo médio de processamento, isso ocorre por seu cálculo ser o mais simples do ponto de vista computacional. O modelo “X” não apresenta uma grande diferença em relação ao modelo “Cruz”, essa diferença ocorre devido ao seu cálculo necessitar mais verificações e mais informações para a obtenção da interatividade do vizinho. O modelo “Anel” possui visivelmente o maior tempo médio de processamento, pois basicamente seu cálculo envolve tanto o modelo “Cruz” quanto o modelo “X”, ou seja, ele analisa todos os oito vizinhos. Entretanto avaliar somente o tempo de processamento não é viável, o interessante é avaliar também a redução do custo energético, apresentado na Tabela 4.3 e 4.4. A partir das tabelas 4.3 e 4.4, é possível verificar que a redução do custo energético entre os modelos e a quantidade de passos de Monte Carlo. A figura 4.1 apresenta de forma mais clara a diferença entre os modelos para a rede de *score* 0,7.

As combinações de modelo e passos de Monte Carlo apresentam uma porcentagem de redução média muito próximas. Por meio de um gráfico é possível verificar, de forma mais consistente, a diferença entre os modelos. Para isso, foi utilizado a rede com *score* 0,7 por apresentar mais interações e proteínas (Figura 4.1).

Tabela 4.3: Comparação entre as combinações dos modelos e passos de Monte Carlo em relação a redução do custo energético em cada processo para a rede de *score* 0,7.

Modelo	Passo de Monte Carlo	Diferença do Custo Energético ($\varepsilon\Delta$)	Desvio Padrão($\varepsilon\Delta$)	Redução do Custo Energético(%)	Desvio Padrão (%)
Cruz	3.000	6,4 bilhões	52,2 milhões	92,06	0,24
	5.000	6,3 bilhões	68,8 milhões	92,18	0,26
	10.000	6,3 bilhões	76,1 milhões	92,54	0,29
X	3.000	6,4 bilhões	81,3 milhões	91,74	0,29
	5.000	6,5 bilhões	57,8 milhões	92,19	0,19
	10.000	6,4 bilhões	47,8 milhões	92,52	0,19
Anel	3.000	12,8 bilhões	121,3 milhões	92,54	0,20
	5.000	12,8 bilhões	111,9 milhões	93,03	0,14
	10.000	12,8 bilhões	116,1 milhões	93,33	0,15

Tabela 4.4: Comparação entre as combinações dos modelos e passos de Monte Carlo em relação a redução do custo energético em cada processo para a rede de *score* 0,8.

Modelo	Passo de Monte Carlo	Diferença de Custo Energético ($\varepsilon\Delta$)	Desvio Padrão ($\varepsilon\Delta$)	Redução de Custo Energético (%)	Desvio Padrão (%)
Cruz	3000	5,5 bilhões	54,7 milhões	92,48	0,41
	5000	5,7 bilhões	60,7 milhões	93,06	0,35
	10000	5,7 bilhões	38,6 milhões	93,19	0,25
X	3000	5,6 bilhões	75,9 milhões	92,44	0,23
	5000	5,7 bilhões	33,4 milhões	93,1	0,16
	10000	5,7 bilhões	58 milhões	93,21	0,25
Anel	3000	11,6 bilhões	100,6 milhões	93,13	0,06
	5000	11,6 bilhões	84,2 milhões	93,58	0,22
	10000	11,3 bilhões	123,3 milhões	93,9	0,11

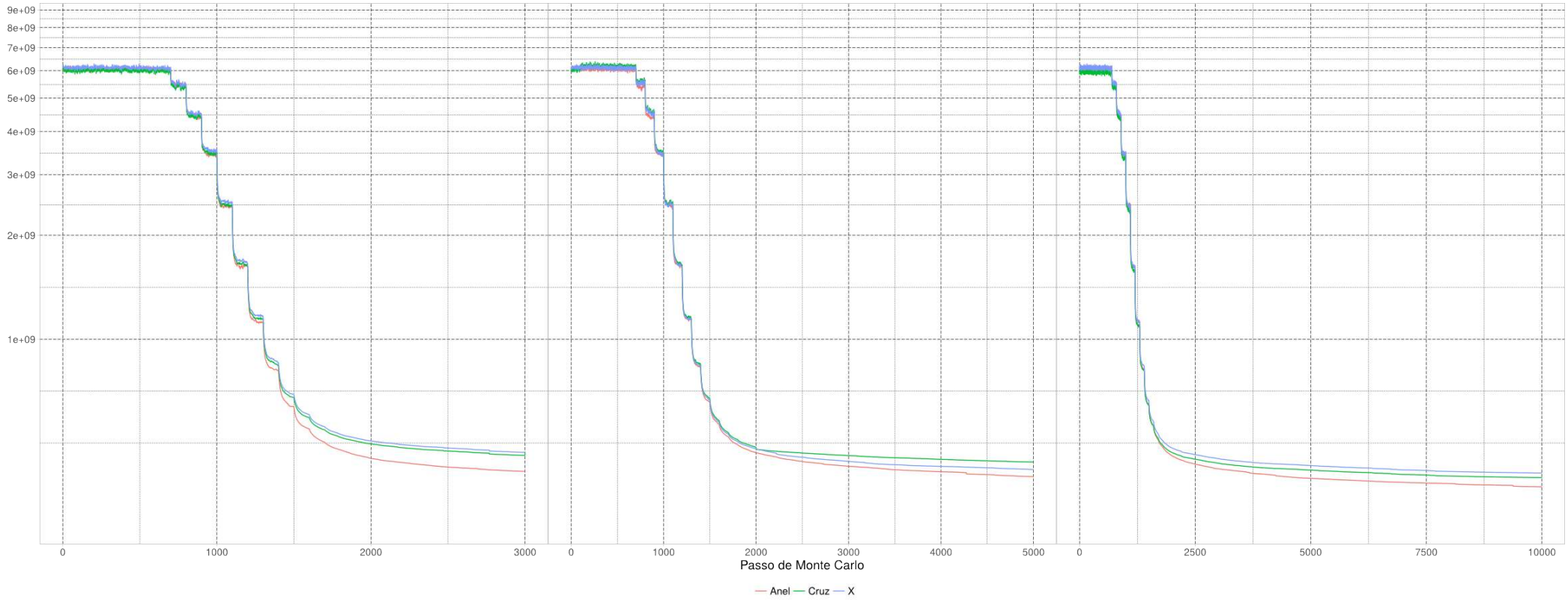

Figura 4.1: Comparação da redução de custo energético, em $\log_2$, em função do passo de Monte Carlo para os três modelos “Anel” em Vermelho,“Cruz” em verde e “X” em azul. Cada gráfico apresenta 3.000, 5.000 e 10.000 passos de Monte Carlos.

Conforme os gráficos da Figura 4.1, a diferença do custo energético entre os modelos não é tão vísivel durante os primeiros 1.000 passos de Monte Carlo. Com a evolução dos passos de Monte Carlo, a diferença torna-se mais perceptível a partir de 2.000 passos de Monte Carlo. Verificando todos os resultados, o modelo “Anel” apresentou, nas três quantidades de passos de Monte Carlo, a melhor redução de custo energético. Em relação a quantidade de passos de Monte Carlo quanto maior a quantidade de passos maior será a redução, em contrapartida o tempo de processamento aumenta.

Na figura 4.2, na parte da evolução da matriz adjacente, verifica-se que a partir de 3.200 passos de Monte Carlo, a clusterização não sofre tanta alteração. Entretanto ao avaliar essa mesma quantidade de passos de Monte Carlo no gráfico de custo energético, verifica-se que ainda pode haver uma redução considerável do custo energético até aproximadamente 6.400 passos de Monte Carlo. A partir de 6.400 a redução do custo energético sofre um decaimento muito baixo, não tornando eficiente o tempo gasto para essa redução. Com isso recomenda-se ao menos utilizar 5.000 passos de Monte Carlo junto com o modelo “Anel”.

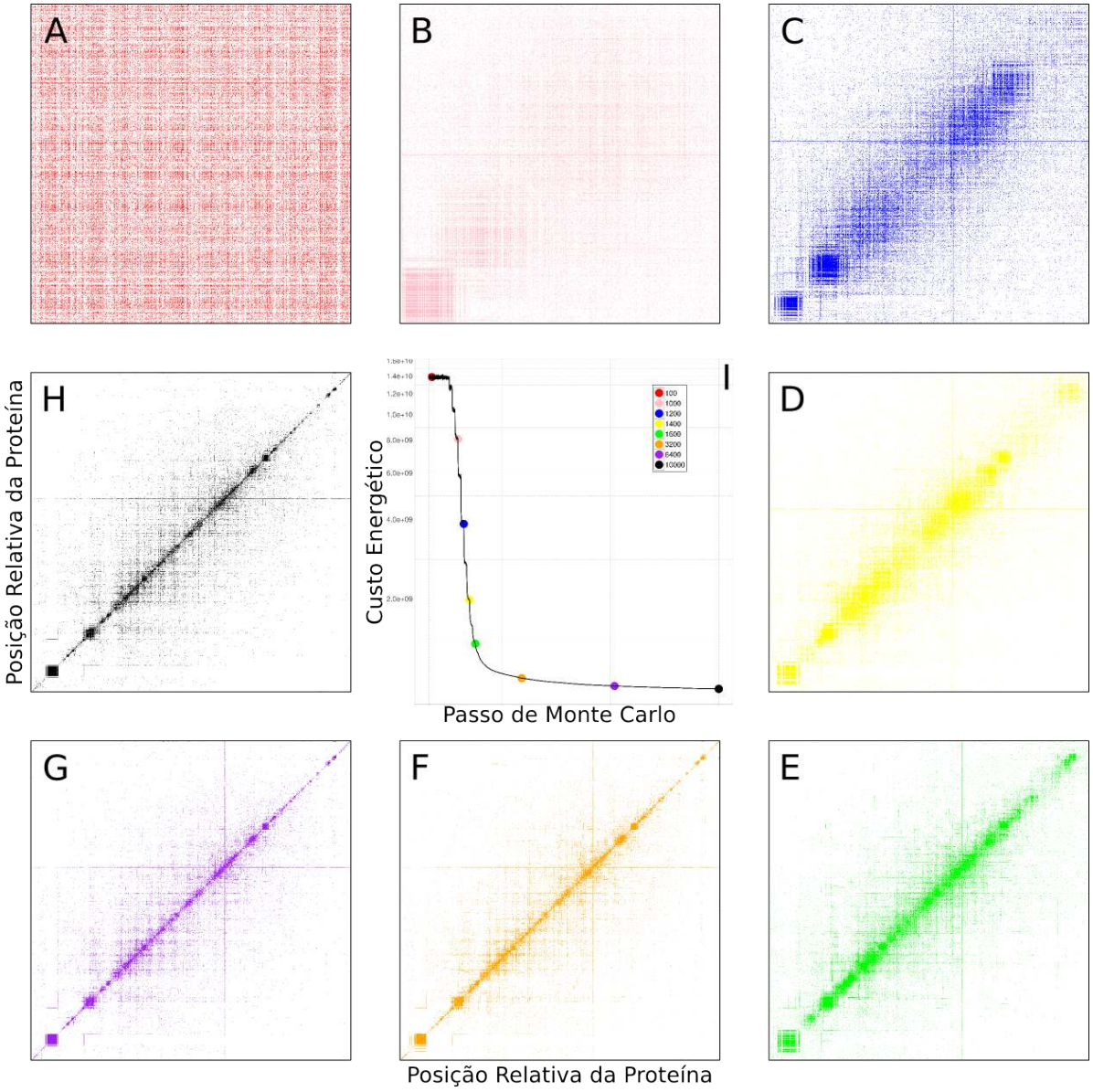

Figura 4.2: Evolução da matriz adjacente ao longo dos passos de Monte Carlo, em sentido horário, a partir de (A) 0 passo de Monte Carlo;(B) 1.000 passos de Monte Carlo; (C) 1.200 passos de Monte Carlo; (D) 1.400 passos de Monte Carlo; (E) 1.600 passos de Monte Carlo; (F) 3.200 passos de Monte Carlo; (G) 6.400 passos de Monte Carlo; (H) 10.000 passos de Monte Carlo; Ao centro, (I) gráfico do custo energético em função dos Passos de Monte Carlo, onde apresenta o custo energético referente a cada quantidade de passos de Monte Carlo apresentado anteriormente.

4.2 Comparação entre as diferentes metodologias de ordenamento

Avaliando os resultados dos tempos de processamento dos modelos de ordenamento para os três modelos com 10.000 passos de Monte Carlo para à rede de *score* 0,7, o resultado também

foi comparado com as primeiras versões da metodologia de ordenamento.

Tabela 4.5: Comparação entre o tempo médio de processamento da metodologia criada por (RYBARCZYK-FILHO et al., 2011), (MOLAN; RYBARCZYK-FILHO, 2014) e Biazotti nos modelos “Cruz”, “X” e “Anel”

Autor	**Modelo**	Passo de Monte Carlo	N. Elementos	N. Interações	Tempo (horas)
(RYBARCZYK-FILHO et al., 2011)	Cruz	2.500	9.019	111.602	720
(MOLAN; RYBARCZYK-FILHO, 2014)		2.500	1.000	199.188	40,57
(KUENTZER, 2014)		10.000	8.815	138.568	0,58
Biazotti		10.000	11.350	463.208	0,86
(MOLAN; RYBARCZYK-FILHO, 2014)	X	2.500	1.000	199.188	49,69
Biazotti		10.000	11.350	463.208	0,96
(MOLAN; RYBARCZYK-FILHO, 2014)	Anel	2.500	1.000	199.188	-
Biazotti		10.000	11.350	463.208	1,19

Na tabela 4.5 verifica-se que ao longo das versões houve uma grandereдução de tempo de processamento, passando de dias para horas , mesmo contendo um número de elementos e interações muito superiores. Logo, concluí-se que a primeira técnica utilizada por Rybarczyk-Filho *et al.* (RYBARCZYK-FILHO et al., 2011) era promissora. Porém, por realizar uma massiva quantidade de trocas deixava o algoritmo lento, precisando de muitos dias para realizar seu ordenamento. A técnica utilizada por Molan *et al.* (MOLAN; RYBARCZYK-FILHO, 2014) resolveu o problema da massiva quantidade de trocas, porém dificultou a procura por vizinhos. A técnica proposta de matriz adjacente e adição de outras técnicas proposta por Kuentzer *et al* (KUENTZER, 2014) trouxe um grande aumento na performance. Ao se verificar de forma imediata o tempo gasto no processamento dos método, é possível interpretar de forma incorreta que o método apresentado nesse trabalho não foi o mais eficiente, entretanto o número de interações e de elementos são inferiores quando comparados aos analisados nesse trabalho, porém conforme aumentam as quantidades de interação e porteínas, mais tempo é necessário para o processamento das informações. Então, a adição dos dois novos melhoramentos, os arquivos de verificação do andamento do ordenamento e o dicionário, foi de grande valia, pois reduziu o tempo de processamento dos dados sem haver perda na redução do custo energético.

4.3 Resultados das modularidades

A partir do ordenamento da rede, foi aplicada uma sequência de janelas de modularidade em cada modelo de ordenamento, isto resultou um vários perfis de modularidade para cada modelo. Na Figura 4.3, verifica-se que conforme aumenta a janela, diminui o número de módulos interativos, e aumenta a interação em um único módulo. Isso se repete independentemente do modelo utilizado como demonstrado nas Figuras 4.4 e 4.5.

A partir da obtenção dos perfis de modularidade dos três modelos, verificou-se que a melhor

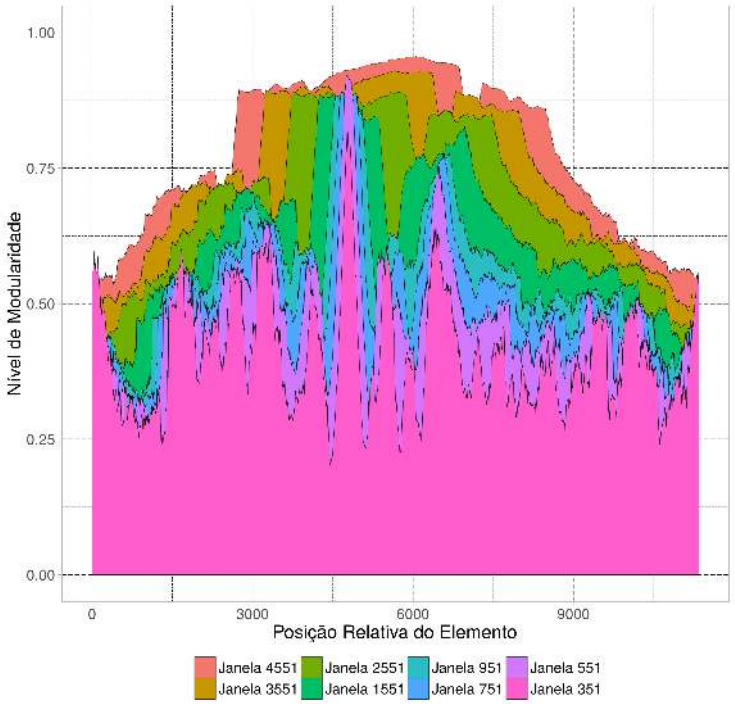

Figura 4.3: Múltiplos perfis de modularidade obtido pelo ordenamento da rede utilizando o modelo “Cruz”.

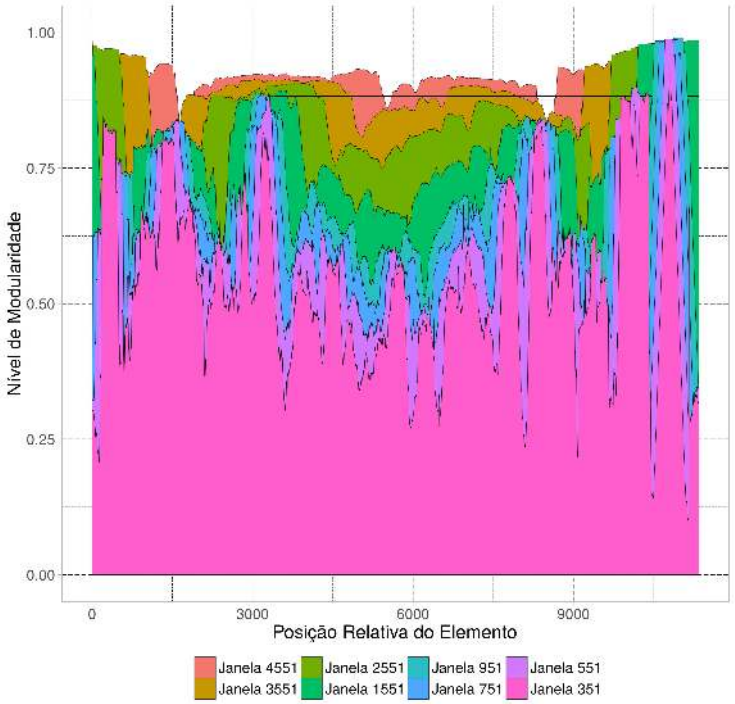

Figura 4.4: Múltiplos perfis de modularidade obtido pelo ordenamento da rede utilizando o modelo “X”.

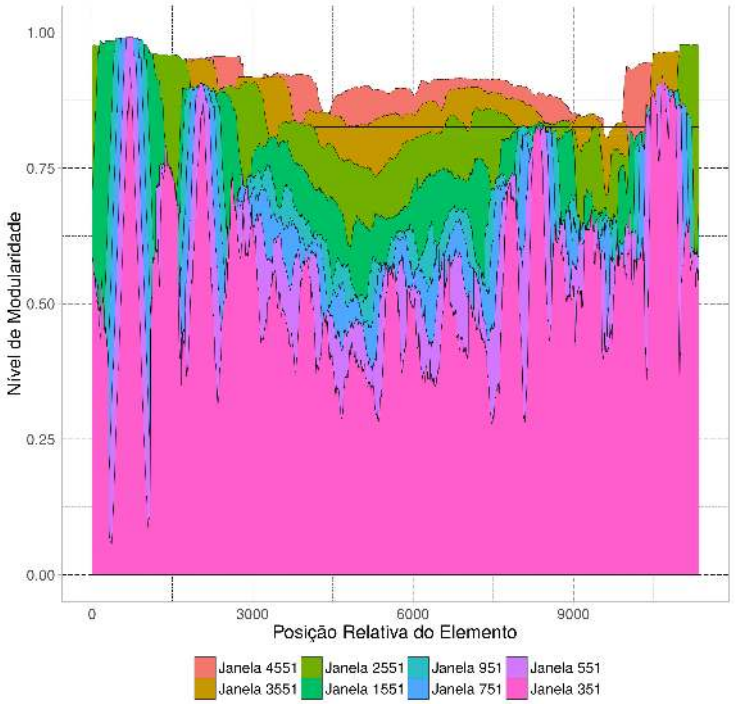

Figura 4.5: Múltiplos perfis de modularidade obtido pelo ordenamento da rede utilizando o modelo “Anel”.

janela é a 351 independente do modelo utilizado para a rede em questão (Figura 4.6). Representada pela cor rosa nas figuras 4.3, 4.4 e 4.5, por apresentar uma quantidade representativa de módulos interativos, representados pelos picos, e cada um deles tendo um valor médio de interação. Comparando os três resultados obtidos com os perfis de modularidade, verificou-se que o modelo “Anel” apresenta uma melhor formação de módulos interativos, onde cada módulo é bem evidenciado, ou seja, cada módulo do método “Anel” apresenta de forma mais concisa os processos biológicos presentes.

Os resultados obtidos, o modelo “Anel” apresentaram o melhor resultado tanto para o ordenamento quanto para a modularidade, sendo assim as próximas análises foram realizadas para o modelo “Anel”.

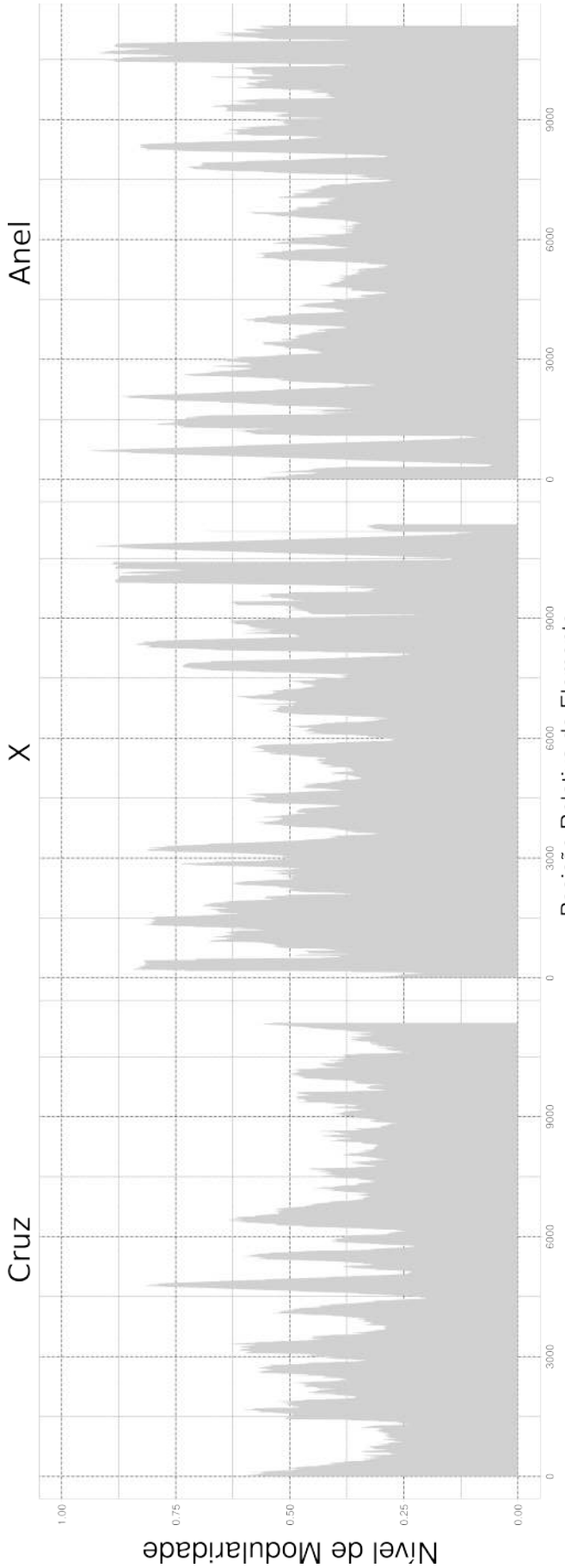

Figura 4.6: Perfis de modularidade obtidos com a janela 351 para cada modelo.

4.4 Análise de Expressão

Os conjuntos de expressão gênica foram prospectados no Gene Expression Omnibus sob os códigos GSE10072 (LANDI et al., 2008) e GSE19804 (LU et al., 2010). Estes dados foram normalizados utilizando a técnica RMA. Além da normalização foi aplicado a suavização dos dados de expressão e obtidos seu respectivo p-valor para cada sonda.

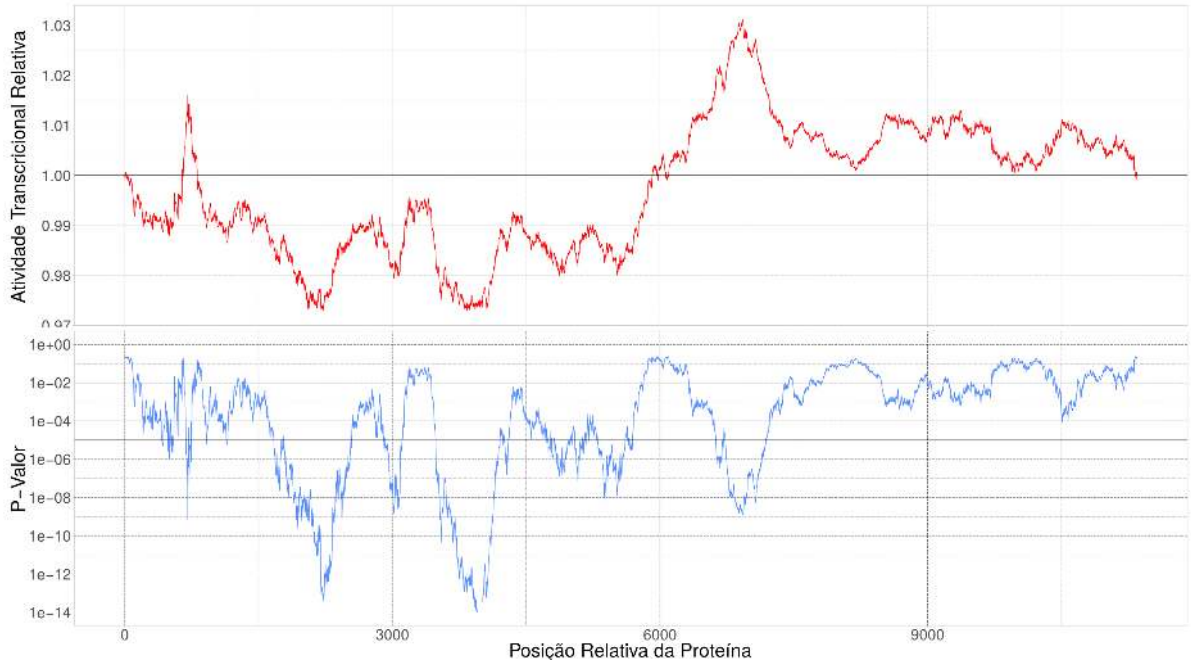

Figura 4.7: Exemplo de resultado da análise do Transcriptograma para a série GSE19804 utilizando a técnica RMA e seu respectivo p-valor, em azul. O Cálculo da expressão obtida da comparação entre taiwanesas não fumantes com câncer e sem câncer de pulmão, em vermelho. A linha preta representa o *cut-off* para o p-valor maior e menor a 1×10^{-5}.

Após a obtenção do p-valor referente a cada comparação, é possível verificar que existe uma certa quantidade de sondas com p-valor menor e igual à 1×10^{-5} para cada comparação, sendo assim nas comparações da séries GSE10072 (amostras de indivíduos italianos masculino e feminino) foi tomado como referência o grupo de italianos não-fumantes sem câncer: 1326 genes alterados no grupo de italianos fumantes com câncer, 353 genes alterados em italianos ex-fumantes com câncer, 387 genes alterados em italianos não-fumantes com câncer, e nenhum gene alterado nos grupos de italianos fumantes sem câncer e italianos ex-fumantes sem câncer.

Na série GSE19804 (amostras de taiwanesas não-fumantes), foi tomado como referência o grupo de taiwanesas sem câncer e como teste o grupo de taiwanesas com câncer, onde foi encontrado na comparação 748 proteínas com p-valor menor ou igual a 1×10^{-5}.

Na Figura 4.8, verifica-se que existem 172 geness em comum entre as 4 comparações com

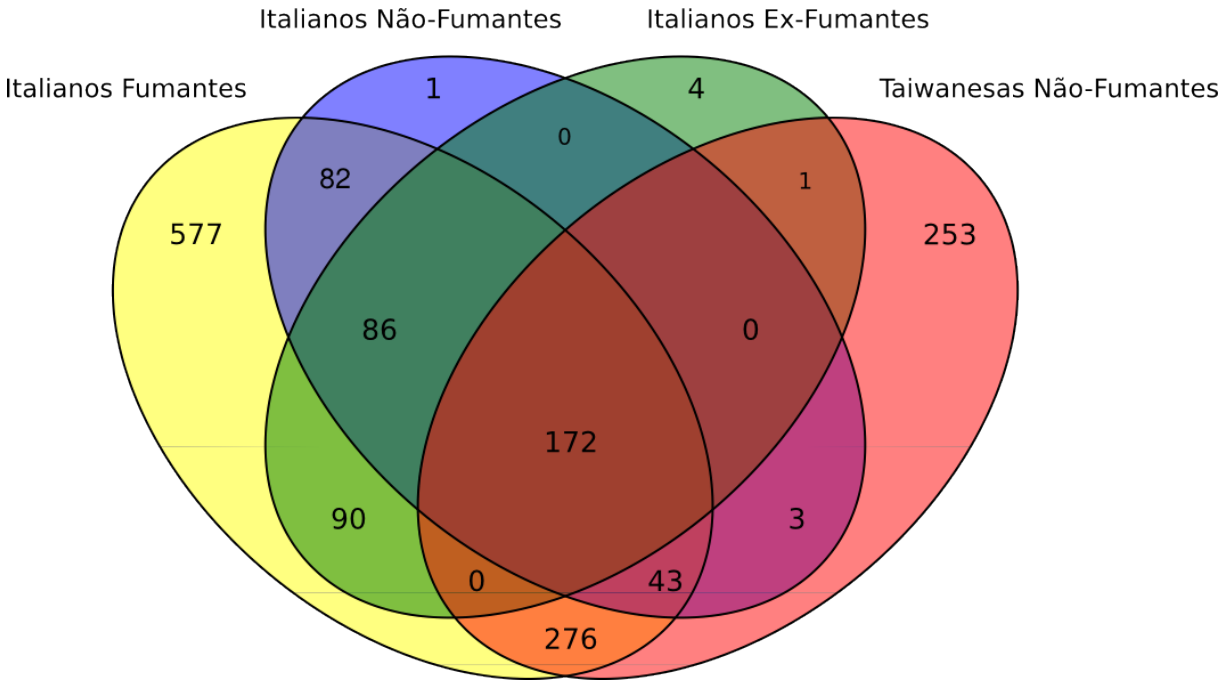

Figura 4.8: Diagrama de Venn para as proteínas obtidas pelo corte de 1×10^{-5} nas comparações de indivíduos com câncer.

câncer, além de outras proteínas que se apresentam comumente entre dois ou três grupos, entretanto cada comparação apresentou uma quantidade distinta de genes que não são compartilhadas com nenhuma outra, sendo 577 genes da comparação de italianos fumantes com câncer, 1 gene da comparação de italianos não-fumantes com câncer, 4 genes da comparação de italianos ex-fumantes com câncer e 253 genes da comparação de taiwanesas não-fumantes com câncer.

4.5 Enriquecimento Funcional

Utilizando o perfil de modularidade obtido a partir do modelo “Anel” (Figura 4.6), realiza-se a separação dos módulos para a aplicação do enriquecimento funcional. Verifica-se quais proteínas, obtidas anteriormente, estão presentes em cada módulo do perfil para obter as ontologias presentes.

A separação dos módulos é realizada de forma manual, é considerado como módulo um grupo de proteínas separadas por dois grandes vales, conforme a Figura 4.9, por meio deles são realizados os enriquecimentos funcionais retornando uma grande quantidade de processos biológicos.

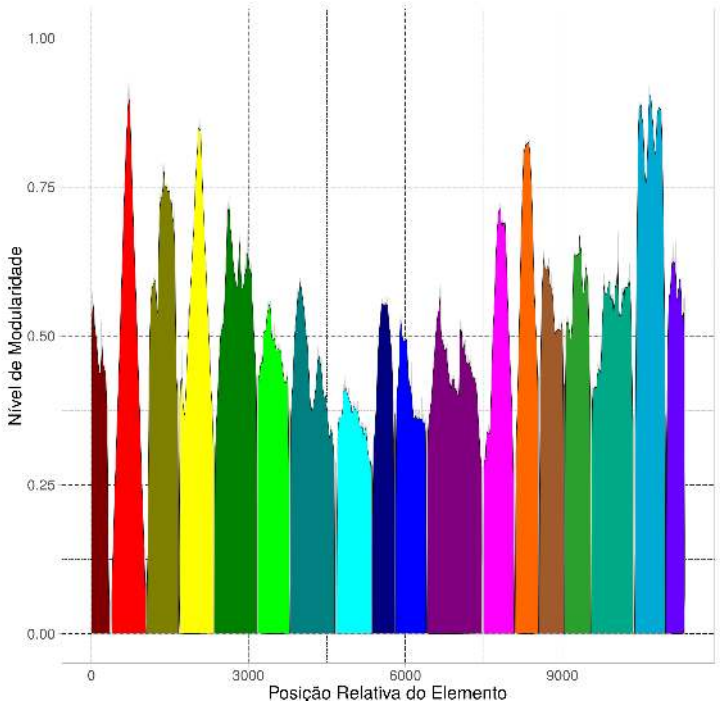

Figura 4.9: Perfil de modularidade referente ao modelo “Anel” com janela 351, foram separadas em 18 módulos, sendo cada módulo representado por uma cor distinta.

As Tabelas 4.6, 4.7, 4.8 e 4.9 apresentam os três processos biológicos mais representativo de cada grupo. Nas tabelas são apresentados apenas os processos com a maior razão de proteínas em relação ao número total de proteínas presentes no processo biológico.

Tabela 4.6: Processos biológicos com maiores razões obtidos a partir das proteínas com p-valor menor que 1×10^{-5} do grupo de pessoas fumantes com câncer da série GSE10072.

GO	Descrição	p-valor	p-ajustado	q-valor	Módulo	Razão
GO:0002003	Maturação da angiotensina	7,57484704577136E-012	5,99394841232982E-010	5,10453143025957E-010	Grupo 3	0,75
GO:0035095	Resposta comportamental à nicotina	3,01430352923168E-008	1,41539769015461E-006	1,20537273574944E-006	Grupo 3	0,75
GO:0036376	Exportação de ion de sódio da célula	1,77549205984646E-010	1,26444626195399E-008	1,07682035980512E-008	Grupo 3	0,73
GO:0060406	Regulação positiva da ereção peniana	4,42921135658892E-012	5,19108736253278E-011	3,37509031714294E-011	Grupo 4	1
GO:0060158	Via de sinalização do receptor de dopamina activadora de fosfolipase C	9,23815402359545E-013	1,1426885899955E-011	7,42942070950203E-012	Grupo 4	0,89
GO:0060405	Regulação da ereção peniana	9,23815402359545E-013	1,1426885899955E-011	7,42942070950203E-012	Grupo 4	0,89
GO:0052697	Glucuronidação xenobiótica	2,43778099016837E-013	6,57760304515911E-012	5,33622065825061E-012	Grupo 5	1
GO:1904223	Regulação da atividade da glucuronosiltransferase	6,17582209870198E-012	1,51986303187287E-010	1,2330212773117E-010	Grupo 5	1
GO:1904224	Regulação negativa da atividade da glucuronosiltransferase	6,17582209870198E-012	1,51986303187287E-010	1,2330212773117E-010	Grupo 5	1
GO:0035766	Quimiotaxia celular ao factor de crescimento dos fibroblastos	9,7954723945243E-007	5,78629501389171E-006	3,10228619262835E-006	Grupo 6	0,67
GO:0035768	Quimiotaxia das células endoteliais ao factor de crescimento dos fibroblastos	9,7954723945243E-007	5,78629501389171E-006	3,10228619262835E-006	Grupo 6	0,67
GO:0050861	Regulação positiva da via de sinalização do receptor de células B	9,7954723945243E-007	5,78629501389171E-006	3,10228619262835E-006	Grupo 6	0,67
GO:0090080	Regulação positiva da cascata MAPKKK pela via de sinalização do receptor do fator de crescimento de fibroblastos	0,000000892	5,46629339623761E-006	2,95308475576761E-006	Grupo 7	0,67
GO:2000544	Regulação da quimiotaxia das células endoteliais ao factor de crescimento dos fibroblastos	0,000000892	5,46629339623761E-006	2,95308475576761E-006	Grupo 7	0,67
GO:2000251	Regulação positiva da reorganização do citoesqueleto de actina	2,49330226796795E-013	2,79920321849175E-012	1,51222844322868E-012	Grupo 7	0,6
GO:0033139	Regulação da fosforilação peptidil-serina da proteína STAT	2,5351182770369E-027	7,39045738775922E-026	4,40928553587108E-026	Grupo 8	0,89
GO:0033141	Regulação positiva da fosforilação peptidil-serina da proteína STAT	2,5351182770369E-027	7,39045738775922E-026	4,40928553587108E-026	Grupo 8	0,89
GO:0060397	Cascata JAK-STAT envolvida na via de sinalização da hormona do crescimento	4,3817182130959E-033	1,7534839612771E-031	1,04616142982529E-031	Grupo 8	0,87
GO:1902262	Processo apoptótico envolvido na padronização dos vasos sanguíneos	9,9256545120455E-007	6,8875880744794E-006	3,69359384543091E-006	Grupo 9	0,67
GO:0044332	Wnt sinalizando caminho envolvido no dorsal / especificação do eixo ventral	2,22262356496031E-008	2,04625343591892E-007	1,09734046452441E-007	Grupo 9	0,71
GO:0003306	Wnt sinalizando caminho envolvido no desenvolvimento do coração	8,61091769917441E-011	1,16902846853269E-009	6,26912688420648E-010	Grupo 9	0,64
GO:0006538	Processo catabólico de glutamato	9,39601767367053E-011	1,66275401868717E-009	1,07392385431218E-009	Grupo 10	1
GO:0009136	Processo de biossíntese de nucleósido difosfato de purina	2,55388410510117E-009	0,000000038	2,45680663911779E-008	Grupo 10	1
GO:0009180	Processo de biossíntese de ribonucleósido difosfato de purina	2,55388410510117E-009	0,000000038	2,45680663911779E-008	Grupo 10	1
GO:0006297	Reparo de nucleotídeo-excisão, preenchimento de fosso de DNA	1,93182523189903E-022	2,12655321527445E-021	1,55684778688621E-021	Grupo 11	1
GO:0031848	Proteção contra a adesão não homóloga no telómero	2,23861548505194E-008	1,51740635834062E-007	1,1108918948378E-007	Grupo 11	1
GO:0045002	Reparo de ruptura de fita dupla através de alinhamento de fita simples	2,23861548505194E-008	1,51740635834062E-007	1,1108918948378E-007	Grupo 11	1
GO:0016246	Interferência de RNA	2,82482941236486E-011	7,28326300376714E-010	6,09568452141892E-010	Grupo 12	0,7
GO:0034475	Processamento de U4 snRNA final-3'	6,96051872328123E-017	1,65505667420243E-015	1,48877761581293E-015	Grupo 13	1
GO:0043634	Processo catabólico de ncRNA dependente de poliadenilação	7,87944857464113E-013	1,28472532950339E-011	1,15565245761403E-011	Grupo 13	1
GO:0071029	Nuclear ncRNA vigilância	7,87944857464113E-013	1,28472532950339E-011	1,15565245761403E-011	Grupo 13	1
GO:0006002	Processo metabólico de Frutose 6-fosfato	4,02856420765214E-012	1,03753197488304E-010	9,03916068752973E-011	Grupo 14	1
GO:0006086	Acetil-CoA biossintético a partir de piruvato	1,366839890454E-014	4,61269186019879E-013	4,01865811712429E-013	Grupo 14	0,77
GO:0010510	Regulação do processo biossintético de acetil-CoA a partir de piruvato	4,52530461120406E-013	1,315910616495E-011	1,14644442782262E-011	Grupo 14	0,75
GO:0016338	Adesão de célula-célula independente do cálcio através de moléculas de adesão celular da membrana plasmática	4,26995506222858E-024	4,24775129590499E-022	3,82947548738816E-022	Grupo 15	0,71
GO:0033572	Transferência de transferrina	1,39693683573944E-034	5,79030318413996E-032	5,22013238618421E-032	Grupo 15	0,69
GO:0015682	Transporte de ferro férrico	1,21298343469547E-033	3,77086225260953E-031	3,39954567881755E-031	Grupo 15	0,66
GO:0019317	Processo catabólico de fucose	4,95050024378372E-016	1,23936207857533E-014	1,15465961364706E-014	Grupo 17	0,89
GO:0042354	Processo metabólico da L-fucose	4,95050024378372E-016	1,23936207857533E-014	1,15465961364706E-014	Grupo 17	0,89
GO:0042355	Processo catabólico de L-fucose	4,95050024378372E-016	1,23936207857533E-014	1,15465961364706E-014	Grupo 17	0,89

Tabela 4.7: Processos biológicos com maiores razões obtidos a partir das proteínas com p-valor menor que 1×10^{-5} do grupo de pessoas ex-fumantes com câncer da série GSE10072.

GO	Descrição	p-valor	p-ajustado	q-valor	Módulo	Razão
GO:0002003	Maturação da angiotensina	7,86596489291291E-023	1,00605690980356E-019	7,17872796016368E-020	Grupo 3	0,75
GO:0002002	Regulação dos níveis de angiotensina no sangue	2,55261792500282E-022	1,6323991630393E-019	1,16479986367234E-019	Grupo 3	0,69
GO:0060177	Regulação do processo metabólico da angiotensina	7,1366479572356E-022	3,04259091243478E-019	2,17104343120114E-019	Grupo 3	0,64
GO:0052697	Glucuronidação xenobiótica	1,92958683354266E-013	5,2738094131488E-012	4,28160178240399E-012	Grupo 5	1
GO:1904223	Regulação da atividade da glucuronosiltransferase	5,01773200126395E-012	1,29213850090352E-010	1,04903724711548E-010	Grupo 5	1
GO:1904224	Regulação negativa da atividade da glucuronosiltransferase	5,01773200126395E-012	1,29213850090352E-010	1,04903724711548E-010	Grupo 5	1
GO:0050861	Regulação positiva da via de sinalização do receptor de células B	1,07544550915807E-007	8,94134834403934E-007	4,7608758287999E-007	Grupo 6	0,67
GO:0090080	Regulação positiva da cascata MAPKKK pela via de sinalização do receptor do fator de crescimento de fibroblastos	9,11765408379687E-008	7,86704187328097E-007	4,27313037873996E-007	Grupo 7	0,67
GO:0033139	Regulação da fosforilação peptidil-serina da proteína STAT	3,97531953535086E-028	1,33032346655442E-026	8,16810371167409E-027	Grupo 8	0,89
GO:0033141	Regulação positiva da fosforilação peptidil-serina da proteína STAT	3,97531953535086E-028	1,33032346655442E-026	8,16810371167409E-027	Grupo 8	0,89
GO:0070669	Resposta à interleucina-2	5,54824305628504E-012	6,51382126773796E-011	3,99944592513249E-011	Grupo 8	0,87
GO:0044332	Wnt sinalizando caminho envolvido na especificação do eixo ventral / dorsal	2,07956306634854E-008	0,000000191	1,0308264769077E-007	Grupo 9	0,71
GO:1902262	Processo apoptótico envolvido na padronização dos vasos sanguíneos	9,41256783931871E-007	6,49310066730544E-006	3,5046235338155E-006	Grupo 9	0,67
GO:0003306	Wnt sinalizando caminho envolvido no desenvolvimento do coração	7,84459942417685E-011	1,0793026608737E-009	5,82549031536482E-010	Grupo 9	0,64
GO:0043401	Via de sinalização mediada pela hormona esteróide	4,86633146397115E-060	9,70493888284146E-058	6,28367948206227E-058	Grupo 10	0,73
GO:2000188	Regulação da homeostase do colesterol	5,12596091110601E-010	1,21248690781931E-008	7,85051734679509E-009	Grupo 10	0,64

Tabela 4.8: Processos biológicos com maiores razões obtidos a partir das proteínas com p-valor menor que 1×10^{-5} do grupo de pessoas não-fumantes com câncer da série GSE10072.

GO	Descrição	p-valor	p-ajustado	q-valor	Módulo	Razão
GO:0060406	Regulação positiva da ereção peniana	5,96924812872284E-011	7,95076921806956E-010	5,0881521445465E-010	Grupo 4	0,86
GO:0007207	Via de sinalização fosfolipase C ativando proteína G acoplado receptor de acetilcolina	3,60655823108529E-009	3,72537487251755E-008	2,38408053698257E-008	Grupo 4	0,83
GO:0060158	Via de sinalização do receptor de dopamina activadora de fosfolipase C	4,28586300416895E-012	6,60209244511765E-011	4,22505670983063E-011	Grupo 4	0,78
GO:0036148	Remodelação de cadeia acilo de fosfatidilglicerol	1,90747236124961E-020	7,98545710845468E-019	6,43296759542536E-019	Grupo 5	0,88
GO:0046341	Processo metabólico do CDP-diacilglicerol	3,0388037567995E-016	1,01576137979221E-014	8,18282529370735E-015	Grupo 5	0,86
GO:0006072	Processo metabólico do glicerol-3-fosfato	1,31621441273206E-008	2,65211053630871E-007	2,13650150615632E-007	Grupo 5	0,86
GO:0050861	Regulação positiva da via de sinalização do receptor de células B	1,63080404935808E-007	1,33951121025225E-006	7,20681194007734E-007	Grupo 6	0,67
GO:0090080	Regulação positiva da cascata MAPKKK pela via de sinalização do receptor do fator de crescimento de fibroblastos	1,63080404935808E-007	1,33951121025225E-006	7,20681194007734E-007	Grupo 6	0,67
GO:0090080	Regulação positiva da cascata MAPKKK pela via de sinalização do receptor do fator de crescimento de fibroblastos	1,40595792758133E-007	1,18915133127206E-006	0,000000648	Grupo 7	0,67
GO:0033139	Regulação da fosforilação peptidil-serina da proteína STAT	3,97531953535086E-028	1,33032346655442E-026	8,16810371167409E-027	Grupo 8	0,89
GO:0033141	Regulação positiva da fosforilação peptidil-serina da proteína STAT	3,97531953535086E-028	1,33032346655442E-026	8,16810371167409E-027	Grupo 8	0,89
GO:0070669	Resposta à interleucina-2	5,54824305628504E-012	6,51382126773796E-011	3,99944592513249E-011	Grupo 8	0,87
GO:0044332	Wnt sinalizando caminho envolvido na especificação do eixo ventral / dorsal	2,22262356496031E-008	2,04625343591892E-007	1,09734046452441E-007	Grupo 9	0,71
GO:1902262	Processo apoptótico envolvido na padronização dos vasos sanguíneos	9,9256545120455E-007	6,8875880744794E-006	3,69359384543091E-006	Grupo 9	0,67
GO:0003306	Wnt sinalizando caminho envolvido no desenvolvimento do coração	8,61091769917441E-011	1,16902846853269E-009	6,26912688420648E-010	Grupo 9	0,64
GO:0043401	Via de sinalização mediada pela hormona esteróide	2,87715411802756E-064	1,0654101699056E-060	6,81431238480211E-061	Grupo 10	0,68
GO:0010871	Regulação negativa do processo biossintético do receptor	2,85943452935212E-007	4,79117016388729E-006	3,06440947418065E-006	Grupo 10	0,67
GO:0010887	Regulação negativa do armazenamento de colesterol	2,85943452935212E-007	4,79117016388729E-006	3,06440947418065E-006	Grupo 10	0,67
GO:0019317	Processo catabólico da fucose	2,23393049309001E-015	7,23958956093985E-014	6,76711108433114E-014	Grupo 17	0,89
GO:0042354	Processo metabólico da L-fucose	2,23393049309001E-015	7,23958956093985E-014	6,76711108433114E-014	Grupo 17	0,89
GO:0042355	Processo catabólico de L-fucose	2,23393049309001E-015	7,23958956093985E-014	6,76711108433114E-014	Grupo 17	0,89

Tabela 4.9: Processos biológicos com maiores razões obtidos a partir das proteínas com p-valor menor que 1×10^{-5} do grupo de taiwanesas com câncer da série GSE19804.

GO	Descrição	p-valor	p-ajustado	q-valor	Módulo	Razão
GO:0036376	Exportação de ion de sódio da célula	6,55054259527747E-014	2,23701029628725E-011	1,75025901273817E-011	Grupo 3	0,64
GO:0007196	Via de sinalização do receptor de glutamato acoplado à proteína G adenilato-ciclase	2,26231222476815E-011	2,15647012289984E-010	1,40823589764422E-010	Grupo 4	1
GO:0060406	Regulação positiva da ereção peniana	2,26231222476815E-011	2,15647012289984E-010	1,40823589764422E-010	Grupo 4	1
GO:0007207	Via de sinalização da Fosfolipase C-ativando proteína G-acoplado receptor de acetilcolina	7,54359931662421E-010	6,48598669243349E-009	4,235532500513E-009	Grupo 4	1
GO:0046341	Processo metabólico do CDP-diacilglicerol	2,86673783588743E-019	1,66223798779243E-017	1,36188598139743E-017	Grupo 5	0,86
GO:0006072	Processo metabólico do glicerol-3-fosfato	4,12691575838089E-010	1,29176115375161E-008	1,05835110226573E-008	Grupo 5	0,86
GO:0016024	Processo de biossíntese de CDP-diacilglicerol	1,27086277452495E-017	6,91544866691501E-016	5,66588699314525E-016	Grupo 5	0,85
GO:0035766	Quimiotaxia celular ao fator de crescimento dos fibroblastos	9,9256545120455E-007	5,87171713139901E-006	3,14960334424455E-006	Grupo 6	0,67
GO:0035768	Quimiotaxia das células endoteliais ao fator de crescimento dos fibroblastos	9,9256545120455E-007	5,87171713139901E-006	3,14960334424455E-006	Grupo 6	0,67
GO:0050861	Regulação positiva da via de sinalização do receptor de células B	9,9256545120455E-007	5,87171713139901E-006	3,14960334424455E-006	Grupo 6	0,67
GO:0035766	Quimiotaxia celular ao factor de crescimento dos fibroblastos	9,04095220539294E-007	5,54917862775114E-006	2,99931968061557E-006	Grupo 7	0,67
GO:0035768	Quimiotaxia das células endoteliais ao fator de crescimento dos fibroblastos	9,04095220539294E-007	5,54917862775114E-006	2,99931968061557E-006	Grupo 7	0,67
GO:0090080	Regulação positiva da cascata MAPKKK pela via de sinalização do receptor do fator de crescimento de fibroblastos	9,04095220539294E-007	5,54917862775114E-006	2,99931968061557E-006	Grupo 7	0,67
GO:0033139	Regulação da fosforilação peptidil-serina da proteína STAT	3,10347037874052E-026	8,78110759923391E-025	5,40927772180863E-025	Grupo 8	0,89
GO:0033141	Regulação positiva da fosforilação peptidil-serina da proteína STAT	3,10347037874052E-026	8,78110759923391E-025	5,40927772180863E-025	Grupo 8	0,89
GO:0060397	Cascata JAK-STAT envolvida na via de sinalização da hormona do crescimento	9,80587263526684E-032	3,89867970636644E-030	2,40163795304766E-030	Grupo 8	0,87
GO:0008635	Ativação da atividade de endopeptidase de tipo cisteína envolvida no processo apoptótico pelo citocromo C	1,51313754845301E-011	1,81576505814361E-010	9,17517842895615E-011	Grupo 9	0,87
GO:0032494	Resposta ao peptidoglicano	2,18711503160298E-010	2,27609829841338E-009	1,1501272103527E-009	Grupo 9	0,7
GO:0048262	Determinação da assimetria dorsal / ventral	7,2816707120834E-009	6,51729643843929E-008	3,29323209678988E-008	Grupo 9	0,67
GO:0043401	Via de sinalização mediada pela hormona esteróide	4,91829684725262E-063	9,45542568884316E-060	6,09092235872917E-060	Grupo 10	0,69
GO:0010871	Regulação negativa do processo biossintético do receptor	5,98714390568515E-007	9,20822732694376E-006	5,93168404634828E-006	Grupo 10	0,67
GO:0010887	Regulação negativa do armazenamento de colesterol	5,98714390568515E-007	9,20822732694376E-006	5,93168404634828E-006	Grupo 10	0,67
GO:0070198	Localização de proteínas no cromossomo, região telomérica	3,97822134440402E-010	2,70326274173709E-009	1,93976338680307E-009	Grupo 11	1
GO:0032201	Manutenção de telômeros via replicação semi-conservadora	4,00793740747947E-037	7,82562462156593E-036	5,61538465551588E-036	Grupo 11	0,96
GO:0006297	Reparo de nucleotídeo-excisão, preenchimento de fosso de DNA	1,21531617368841E-024	1,43101160146136E-023	1,02684207041653E-023	Grupo 11	0,94

O enriquecimento dos grupos de expressão apresentou 339, 84, 83 e 177 processos biológicos alterados dentro de um grupo, sendo respectivamente, da série GSE10072 (indivíduos italianos) fumante com câncer, ex-fumante com câncer, não-fumante com câncer e da série GSE19804 taiwanesas não fumantes com câncer.

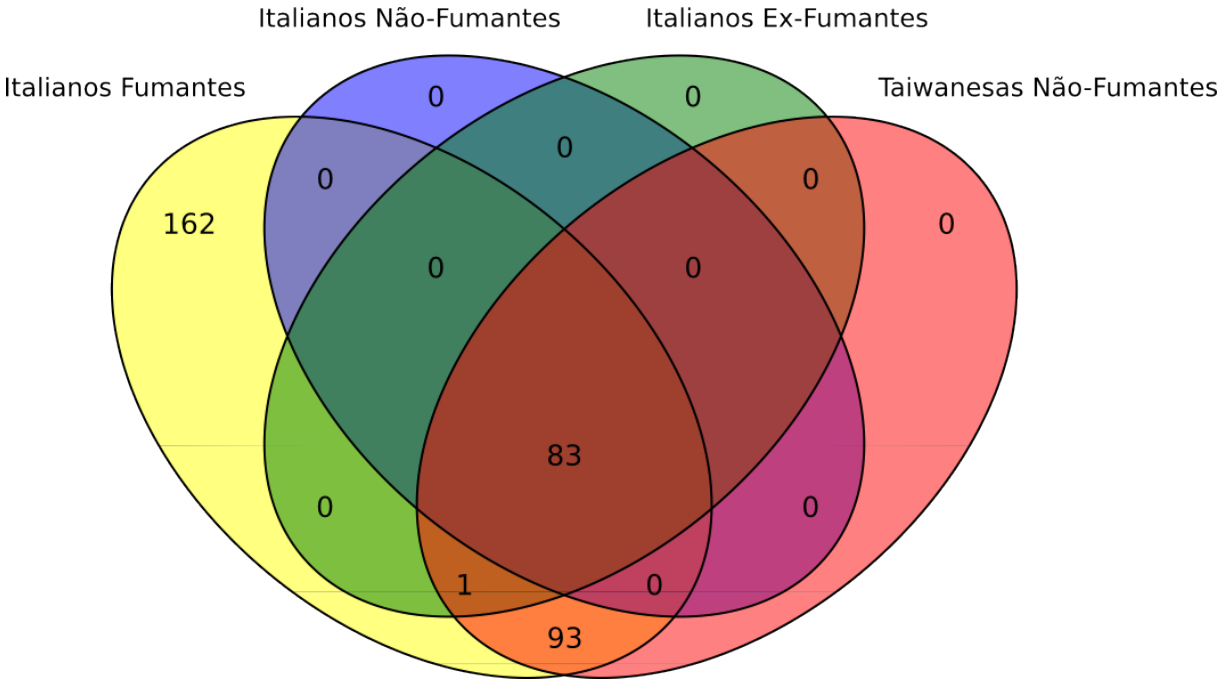

Figura 4.10: Diagrama de Venn entre os processos biológicos obtidos pelo enriquecimento funcional das Comparações de indivíduos com câncer.

O diagrama de Venn (Figura 4.10) apresenta 83 processos biológicos alterados que são comuns as cinco comparações, porém a comparação de italianos fumantes com câncer apresentam 162 processos biológicos que nenhuma outra comparação compartilha. As taiwanesas não-fumantes com câncer compartilham 93 processos biológicos com os italianos fumantes com câncer.

4.6 Transcriptograma

Por meio dos resultados de enriquecimento funcional é possível realizar a integração de todos os resultados em uma única representação gráfica. Esta representação contém o perfil de expressão, os processos biológicos e o p-valor referente a cada comparação caso/teste dos dados de expressão gênica.

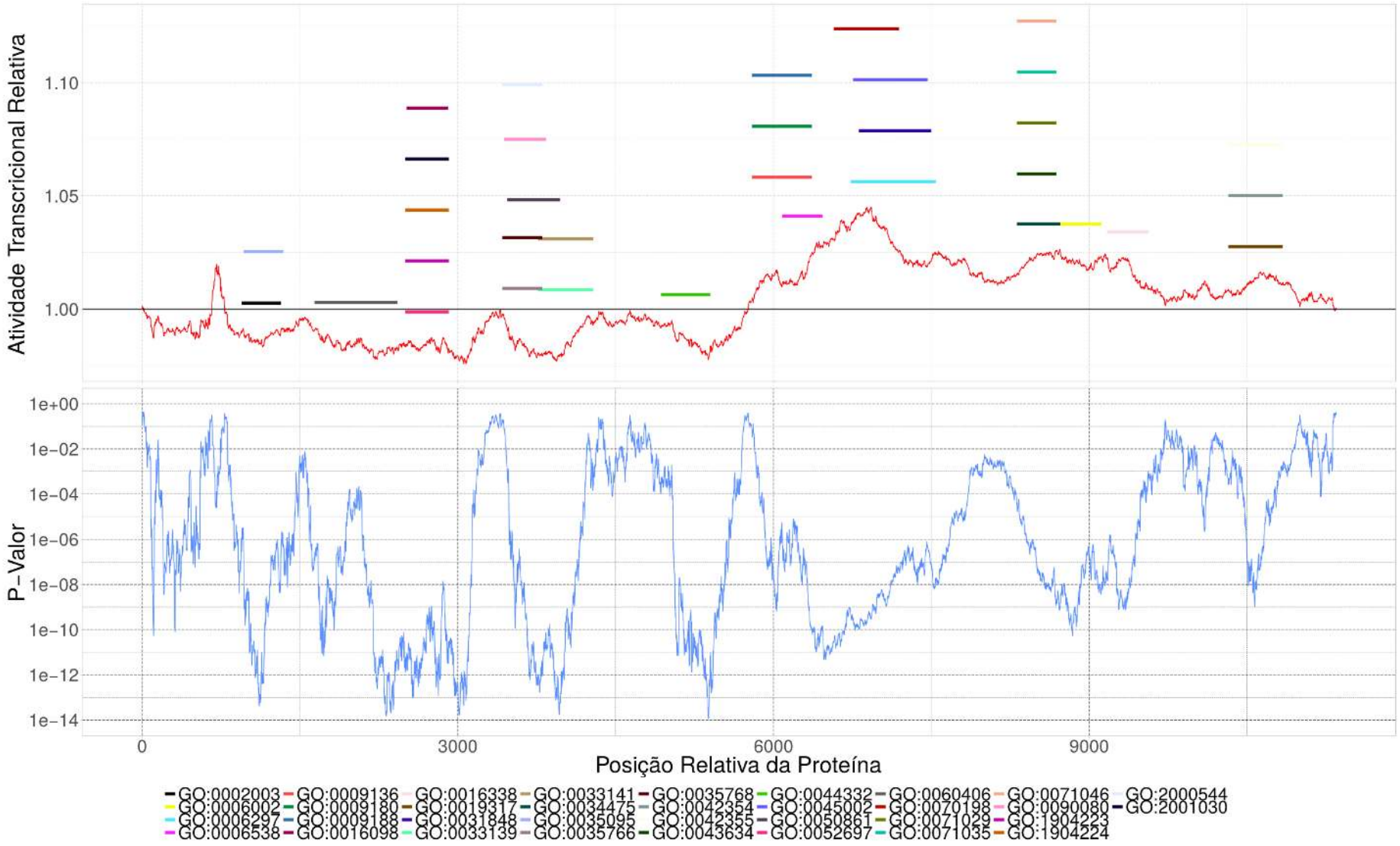

Figura 4.11: Resultado da análise da normalização dos dados da série GSE10072 utilizando a técnica RMA e seu respectivo p-valor, em azul. Perfil de expressão obtido da comparação entre fumante com câncer e não-fumante sem câncer, em vermelho. Linhas de diversas cores representando a posição de alguns processos biológicos que possuem p-valor menor que 1×10^{-5}.

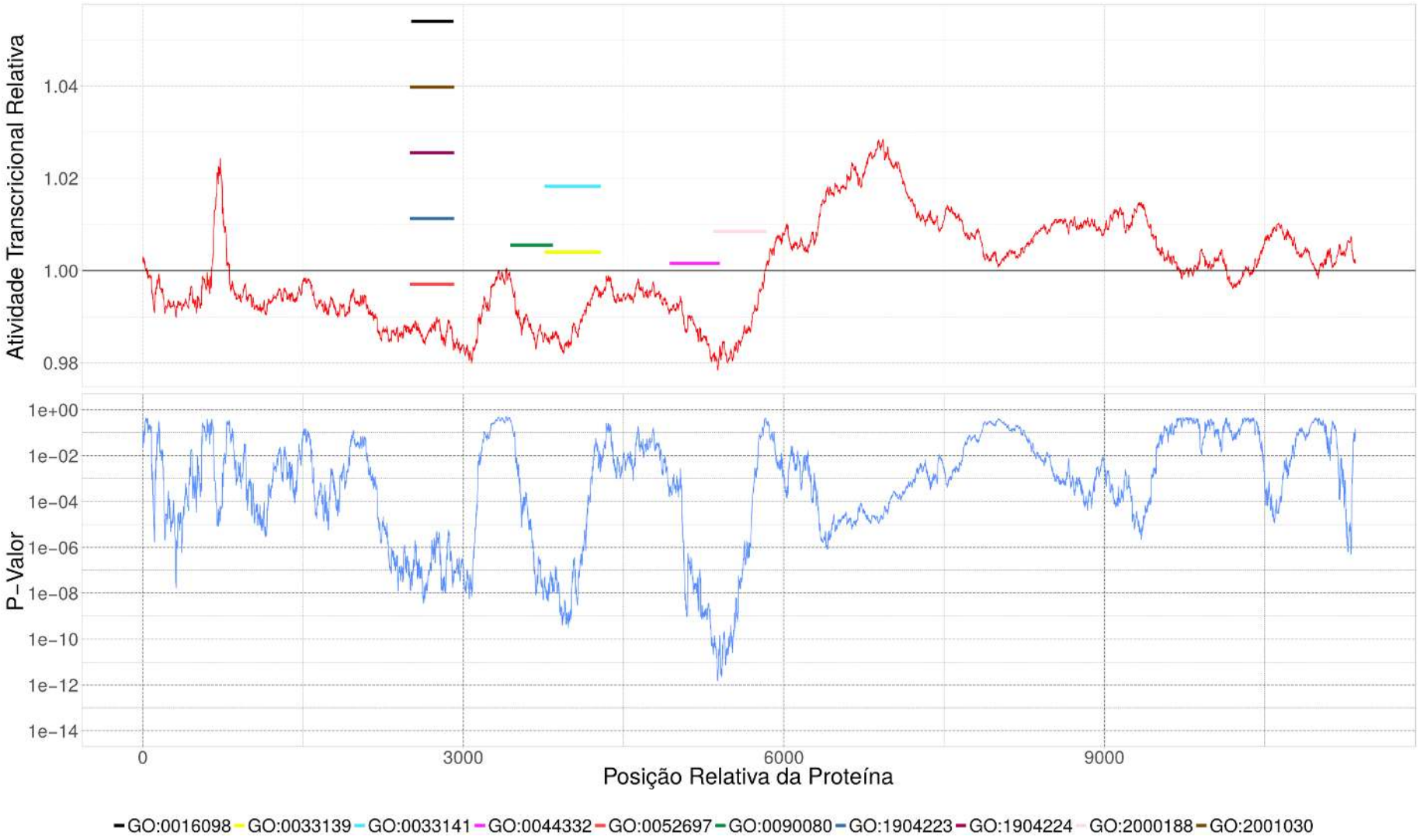

Figura 4.12: Resultado da análise da normalização dos dados da série GSE10072 utilizando a técnica RMA e seu respectivo p-valor, em azul. Perfil de expressão obtido da comparação entre ex-fumante com câncer e não-fumante sem câncer, em vermelho. Linhas de diversas cores representando a posição de alguns processos biológicos que possuem p-valor menor que 1×10^{-5}.

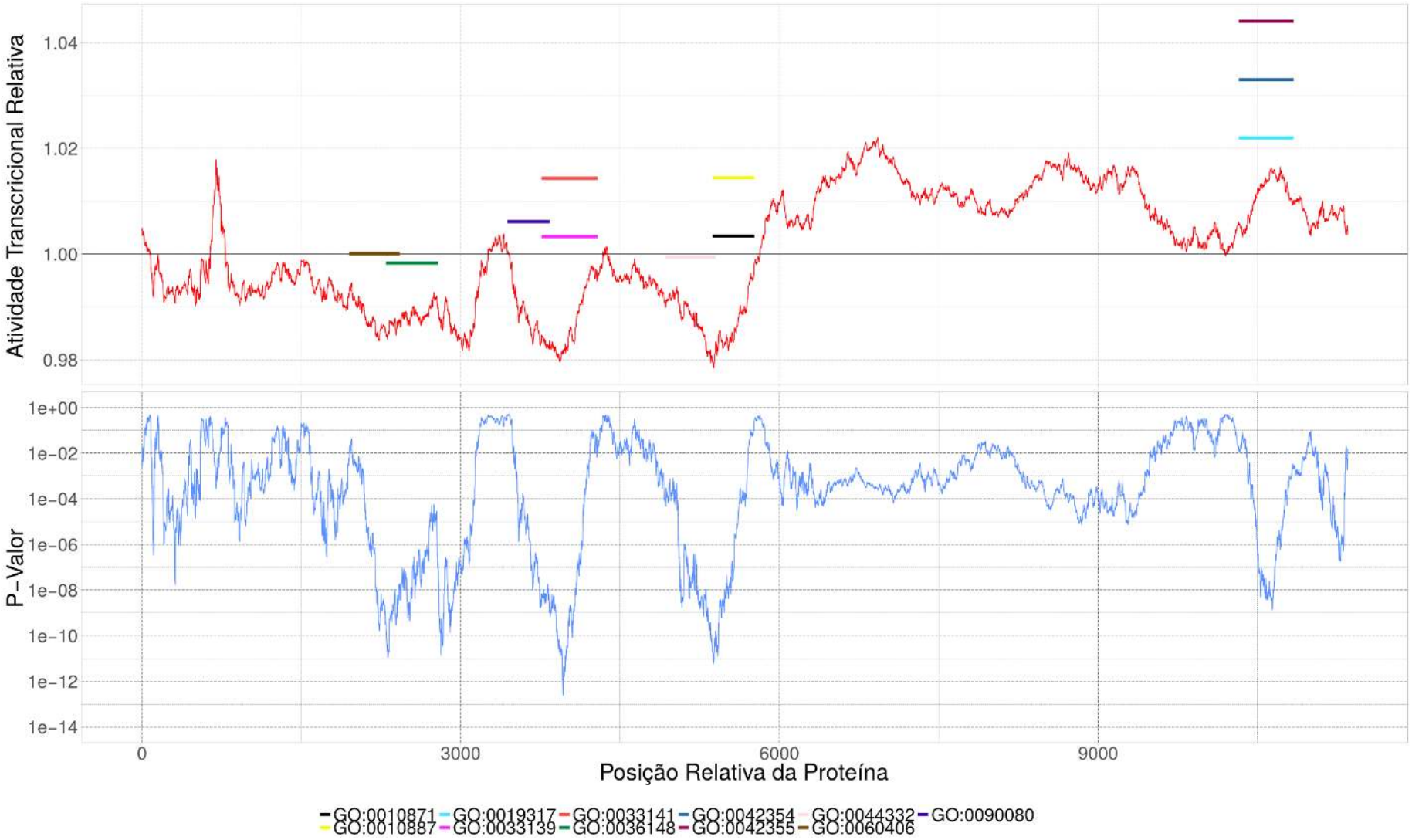

Figura 4.13: Resultado da análise da normalização dos dados da série GSE10072 utilizando a técnica RMA e seu respectivo p-valor, em azul.Perfil de expressão obtido da comparação entre não-fumante com câncer e não-fumante sem câncer, em vermelho. Linhas de diversas cores representando a posição de alguns processos biológicos que possuem p-valor menor que 1×10^{-5}.

Ao comparar os resultados obtidos nas Figuras 4.11, 4.12 e 4.13 com os resultados obtidos pelo artigo de Landi *et al.* (LANDI et al., 2008), é possível observar que os valores de expressão foram analisados de forma mais pontual, ou seja, analisaram a gene à gene. Entretanto, genes que não apresentaram um *fold-change* mínimo de 1,5 para os superexpressos e um máximo de 0,6667 para os subexpressos foram descartados, diferente das análises do transcriptograma que apresentaram todos os genes presentes na rede proteíca. Um segundo ponto, é que por meio de suas análises, foi obtido um p-valor de no mínimo de 1×10^{-3}, enquanto que as análises do transcriptograma foi obtido um p-valor mínimo de 1×10^{-13}, tornando os resultados do transcriptograma mais acurados. Um último ponto, é que ao realizar as comparações foi levado em conta os grupos de uma mesma condição em relação as amostras do grupo de italianos não-fumante e sem câncer, Landi e colaboradores agruparam as amostras de distintas formas, entretanto não diferenciando tanto os tipos das amostras, por exemplo, agruparam todas as amostras de câncer independente se o doador era fumante, ex-fumante ou não fumante e agruparam todos os normais e realizaram a análise, com isso a acurácia e precisão dos dados diminuíram.

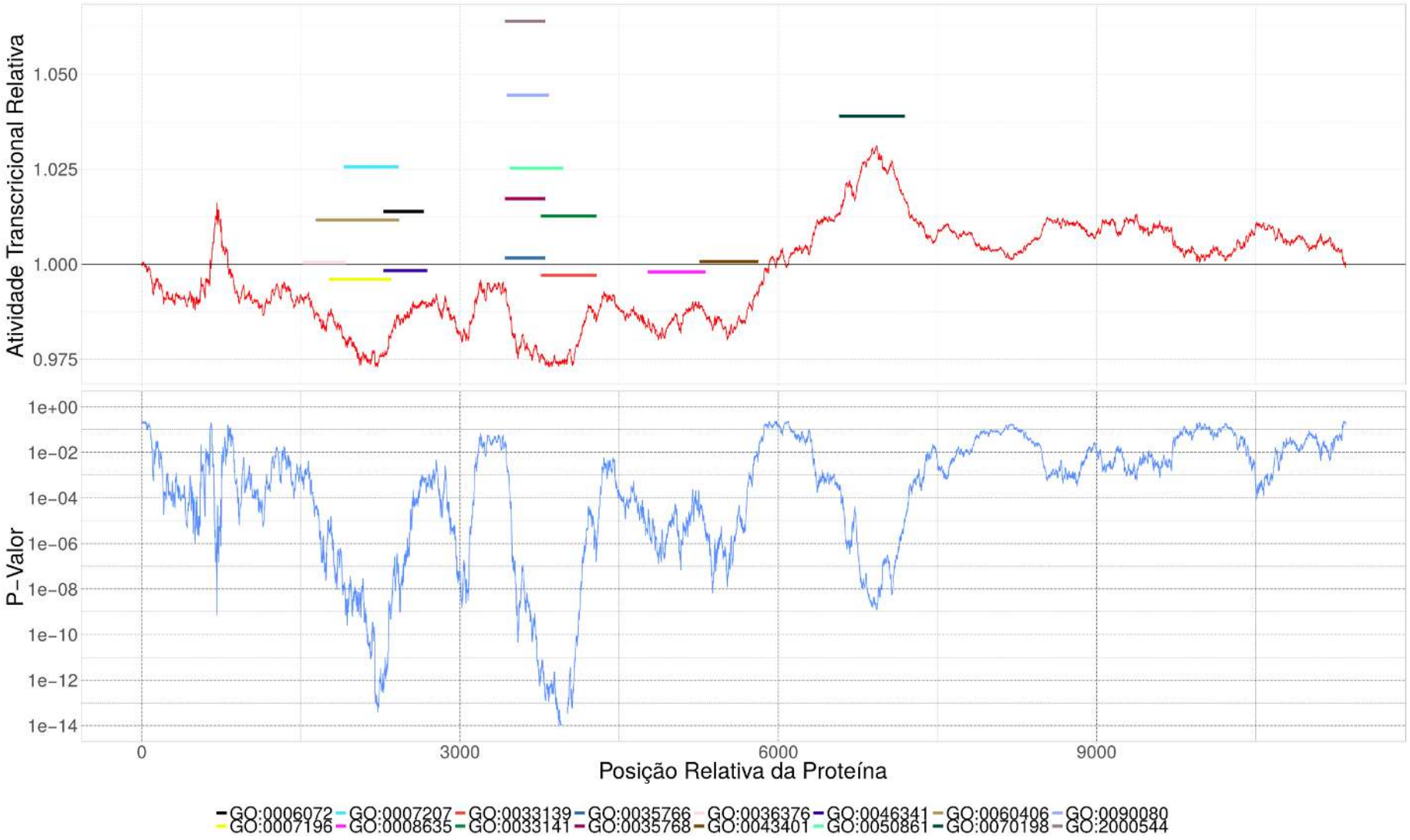

Figura 4.14: Resultado da análise da normalização dos dados da série GSE19804 utilizando a técnica RMA e seu respectivo p-valor, em azul. Perfil de expressão obtido da comparação entre taiwanesas com câncer e taiwanesas sem câncer. Linhas de diversas cores representando a posição de alguns processos biológicos que possuem p-valor menor que 1×10^{-5}.

Ao compararmos o resultado obtido na Figura 4.14 com os resultados obtidos por Lu *et al* (LU et al., 2010), é possível observar que grande parte dos genes foram descartados por não apresentarem um resultado, em log, superior a $-\log(10^{-16})$. Um outro ponto é que o estudo realizado foi direcionado para um único processo biológico, o *Axon guidance*, que vai de encontro com o método do transcriptograma, que visa o estudo da maior quantidade possível de processos biológicos. Um último ponto, os resultados de Lu e colaboradores apresentaram em um dos métodos um p-valor de $4,6 \times 10^{-3}$ e no outro método um q-valor de 1×10^{-4}. Como não é possível comparar diretamente um dos resultados, pois teve como resposta o q-valor. Entretanto, o resultado obtido teve um p-valor de 1×10^{-13} obtido pelo transcriptograma. Numa análise mais profunda, verificamos que o processo biológico *Axon guidance* não está contida nos processos biológicos mais representativos de cada módulo, nem nos restantes dos processos biológicos que apresentam uma razão de no mínimo 70%.

4.7 Comparação entre diferentes metodologias de análise de expressão

Os resultados obtidos pelo transcriptograma são comparados com as metodologias GAGE e PAGE (Figura 4.15).

Por meio da Figura 4.15 é possível verificar que o transcriptograma não possuí nenhum processo em comum com os métodos GAGE e PAGE, devido a utilização da clusterização da matriz e do enriquecimento funcional realizado sobre a rede. Entretanto uma parcela dos processos biológicos presentes no transcriptograma tem processos biológicos mais especifícos (processos biológicos filhas) e processos biológicos mais abrangentes (processos biológicos mães) em relação ao GAGE e PAGE.

Tabela 4.10: Processos biológicos mães e filhas da metodologia transcriptograma em relação as metodologias GAGE e PAGE.

	Processos Biológicos	
Grupo	**Transcriptograma (mãe)**	**Transcriptograma (filha)**
Taiwanesas com Câncer	13	4
Fumante com Câncer	42	23
Ex-Fumante com Câncer	1	0
Não-Fumante com Câncer	1	0

A tabela 4.10 informa que o transcriptograma identificou 57 processos biológicos que são mães dos processos biológicos encontrados pelas metodologias PAGE e GAGE, além disso, encontrou 27 processos biológicos que são filhas dos processos biológicos identificados pelo PAGE e GAGE. A tabela 4.11 informa que a metodologia PAGE identificou 20 processos

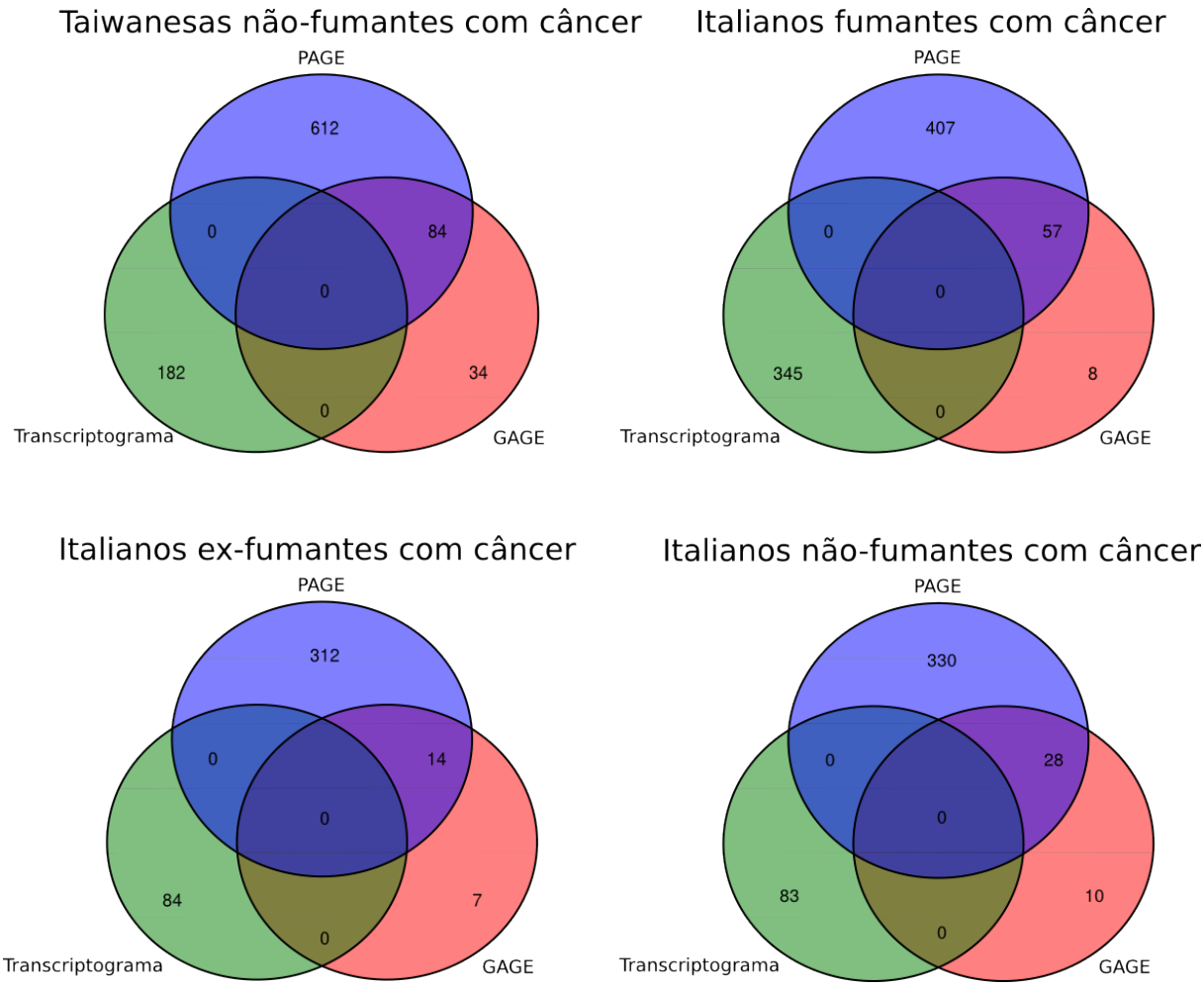

Figura 4.15: Processos biológicos obtidos em três diferentes metodologias de análise de expressão gênica, sendo elas o transcriptograma, GAGE e PAGE, avaliando quatro diferentes situações (comparações) sendo taiwanesas não-fumantes com câncer, italianos fumantes com câncer, italianos ex-fumantes com câncer e italianos não-fumantes com câncer.

biológicos que são mães dos processos biológicos encontrados pela metodologia do transcriptograma, além disso, encontrou 147 processos biológicos filhas. Em relação a metodologia GAGE tanto para processos biológicos mães quanto filhas não foi identificado nenhum processo biológico referente aos do transcriptograma.

Tabela 4.11: Processos biológicos mães e filhas das metodologias GAGE e PAGE em relação a metodologia do transcriptograma.

	Processos Biológicos			
Grupo	**GAGE (mãe)**	**PAGE (mãe)**	**GAGE (filha)**	**PAGE (filha)**
Taiwanesas com Câncer	0	4	0	57
Fumante com Câncer	0	16	0	80
Ex-Fumante com Câncer	0	0	0	7
Não-Fumante com Câncer	0	0	0	3

As análises realizadas pelas metodologias PAGE e GAGE apresentam problemas, pois ambas fornecem uma grande quantidade de falsos positivos, devido ao fato não eliminarem por completo o ruído presente no microarranjo (TRIPATHI; GLAZKO; EMMERT-STREIB, 2013). Quando o ruído não é filtrado de maneira correta, ele pode alterar significativamente o resultado. Estes ruídos presentes no *chip* de microarranjo podem não estar relacionados a expressão de genes e sim a problemas na leitura do *chip*. Ao contrário da metodologia do transcriptograma que ao utilizar integração de vários bancos de dados é capaz de realizar um mensuramento global da expressão (RYBARCZYK-FILHO et al., 2011).

5 *Conclusões*

A alteração realizada no algoritmo de ordenamento, inserindo os arquivos de verificação do processamento dos dados e a alteração da matriz adjacente, teve uma boa performance em comparação aos outros ordenamentos, sendo esse capaz de reduzir de forma drástica o tempo necessário para o ordenamento, quando comparado com os ordenamentos desenvolvidos por (RYBARCZYK-FILHO et al., 2011) e (MOLAN; RYBARCZYK-FILHO, 2014), sem haver perda do poder de redução do custo energético. Ao analisar todos modelos de análise de vizinhança, verificou-se que o modelo "Anel" apresentou os melhores resultados tanto para a redução do custo energético quanto para a clusterização e aproximação da diagonal principal.

Através da automação do processo de enriquecimento funcional, que anteriormente era realizado de forma manual e verificando um-a-um em determinados sites, fez com que reduzisse o tempo necessário para encontrar os processos biológicos referentes as proteínas com um p-valor igual ou inferior a 1×10^{-5} dentro de cada módulo. Além disso aumentamos a acurácia dos resultados por considerar apenas os processos biológicos que apresentavam no mínimo 60% das proteínas dos processos presentes no módulo.

Ao aplicar a metodologia em amostras de câncer em diferentes indivíduos, foi possível verificar uma certa semelhança existente entre esses indivíduos, sendo a superexpressão e subexpressão muito próximos, levando em conta que o nível de expressão das proteínas ainda são diferentes. Além disso muitos processos biológicos são semelhantes nos 4 grupos análisados.

Apesar de alguns processos ainda necessitarem de manipulação manual do usuário, como a seleção dos módulos e a determinação dos parametros, as análises finais apresentam resultados com excelente qualidade. A metodologia apresenta um grande diferencial em relação as outras metodologias de análise de expressão gênica, pois é realizada utilizando uma rede proteíca que permite analisar o organismo de uma forma global. Em comparação aos outros que realizam de forma mais pontual, ou seja, apenas nos genes, além de descartarem genes que não são considerados diferencialmente expressos apresentam uma grande quantidade falsos positivos.

Referências Bibliográficas

BARNES, M. Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms. *Nucleic Acids Research*, Oxford University Press (OUP), v. 33, n. 18, p. 5914?5923, Oct 2005. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gki890>.

BARRETT, T. et al. Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Res*, v. 41, n. Database issue, p. D991–D995, Jan 2013. Disponível em: <http://dx.doi.org/10.1093/nar/gks1193>.

CONSORTIUM, G. O. Gene ontology consortium: going forward. *Nucleic Acids Res*, v. 43, n. Database issue, p. D1049–D1056, Jan 2015. Disponível em: <http://dx.doi.org/10.1093/nar-/gku1179>.

CUI, X.; CHURCHILL, G. A. Statistical tests for differential expression in cdna microarray experiments. *Genome biology*, BioMed Central, v. 4, n. 4, p. 1, 2003.

DALMAN, M. R. et al. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*, v. 13 Suppl 2, p. S11, Mar 2012. Disponível em: <http://dx.doi.org/10.1186/1471-2105-13-S2-S11>.

DU, P.; KIBBE, W. A.; LIN, S. M. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, Oxford University Press (OUP), v. 24, n. 13, p. 1547?1548, May 2008. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btn224>.

DUNNING, M. J. et al. beadarray: R classes and methods for illumina bead-based data. *Bioinformatics*, Oxford University Press (OUP), v. 23, n. 16, p. 2183?2184, Jun 2007. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btm311>.

FAN, J. et al. [3] illumina universal bead arrays. *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*, Elsevier BV, p. 57?73, 2006. ISSN 0076-6879. Disponível em: <http://dx.doi.org/10.1016/S0076-6879(06)10003-8>.

GAUTIER, L. et al. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, Oxford University Press (OUP), v. 20, n. 3, p. 307?315, Feb 2004. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btg405>.

GERSHON, D. Microarray technologyan array of opportunities. *Nature*, Springer Nature, v. 416, n. 6883, p. 885?891, Apr 2002. ISSN 0028-0836. Disponível em: <http://dx.doi.org-/10.1038/416885a>.

GHARAIBEH, R. Z.; FODOR, A. A.; GIBAS, C. J. Background correction using dinucleotide affinities improves the performance of gcrma. *BMC bioinformatics*, BioMed Central, v. 9, n. 1, p. 1, 2008.

GUSNANTO, A.; CALZA, S.; PAWITAN, Y. Identification of differentially expressed genes and false discovery rate in microarray studies. *Current opinion in lipidology*, LWW, v. 18, n. 2, p. 187–193, 2007.

IRIZARRY, R. A. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, Oxford University Press (OUP), v. 31, n. 4, p. 15e?15, Feb 2003. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gng015>.

IRIZARRY, R. A. et al. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, v. 31, n. 4, p. e15, Feb 2003.

IRIZARRY, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, Biometrika Trust, v. 4, n. 2, p. 249–264, 2003.

KIM, C. C. et al. Improved analytical methods for microarray-based genome-composition analysis. *Genome biology*, BioMed Central, v. 3, n. 11, p. 1, 2002.

KIM, S.-Y.; VOLSKY, D. J. Page: parametric analysis of gene set enrichment. *BMC bioinformatics*, BioMed Central Ltd, v. 6, n. 1, p. 144, 2005.

KUENTZER, F. A. *Otimização e análise de algoritmos de ordenamento de redes proteicas*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2014.

LANDI, M. T. et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*, v. 3, n. 2, p. e1651, Feb 2008. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0001651>.

LEUNG, Y. F.; CAVALIERI, D. Fundamentals of cdna microarray data analysis. *Trends in Genetics*, Elsevier BV, v. 19, n. 11, p. 649?659, Nov 2003. ISSN 0168-9525. Disponível em: <http://dx.doi.org/10.1016/j.tig.2003.09.015>.

LOCKHART, D. J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, Springer Nature, v. 14, n. 13, p. 1675?1680, Dec 1996. ISSN 1087-0156. Disponível em: <http://dx.doi.org/10.1038/nbt1296-1675>.

LU, T.-P. et al. Identification of a novel biomarker, sema5a, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*, v. 19, n. 10, p. 2590–2597, Oct 2010. Disponível em: <http://dx.doi.org/10.1158/1055-9965.EPI-10-0332>.

LUO, W. et al. Gage: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, v. 10, p. 161, 2009. Disponível em: <http://dx.doi.org/10.1186/1471-2105-10-161>.

MOLAN, A. L.; RYBARCZYK-FILHO, J. L. Desenvolvimento e comparaï¿½ï¿½o de algoritmos para a organizaï¿½ï¿½o hierï¿½rquica de redes. *Anais do X Congresso de Fï¿½sica Aplicada ï¿½ Medicina*, v. 1, n. 117-121, 2014.

NAEF, F.; MAGNASCO, M. O. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E*, APS, v. 68, n. 1, p. 011906, 2003.

PEPPER, S. D. et al. The utility of mas5 expression summary and detection call algorithms. *BMC bioinformatics*, BioMed Central, v. 8, n. 1, p. 1, 2007.

RYBARCZYK-FILHO, J. L. et al. Towards a genome-wide transcriptogram: the saccharomyces cerevisiae case. *Nucleic Acids Res*, v. 39, n. 8, p. 3005–3016, Apr 2011. Disponível em: <http://dx.doi.org/10.1093/nar/gkq1269>.

STEEMERS, K. L. G. F. J. Illumina, inc. *Pharmacogenomics*, v. 6, n. 7, p. 777–782, Oct 2005.

SZKLARCZYK, D. et al. String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, v. 43, n. Database issue, p. D447–D452, Jan 2015. Disponível em: <http://dx.doi.org/10.1093/nar/gku1003>.

TRIPATHI, S.; GLAZKO, G. V.; EMMERT-STREIB, F. Ensuring the statistical soundness of competitive gene set approaches: gene filtering and genome-scale coverage are essential. *Nucleic Acids Res*, v. 41, n. 7, p. e82, Apr 2013. Disponível em: <http://dx.doi.org/10.1093-/nar/gkt054>.

TSENG, G. C.; GHOSH, D.; FEINGOLD, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, Oxford University Press (OUP), v. 40, n. 9, p. 3785?3799, Jan 2012. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkr1265>.

VERLI, H. et al. Bioinformática da biologia à flexibilidade molecular. *Porto Alegre, Brasil*, v. 1, 2014.

WU, Z. A review of statistical methods for preprocessing oligonucleotide microarrays. *Statistical methods in medical research*, SAGE Publications, v. 18, n. 6, p. 533–541, 2009.

XIE, Y.; PAN, W.; KHODURSKY, A. B. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, Oxford University Press (OUP), v. 21, n. 23, p. 4280?4288, Sep 2005. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bti685>.

YANG, Y. H.; THORNE, N. P. Normalization for two-color cdna microarray data. *Lecture Notes-Monograph Series*, JSTOR, p. 403–418, 2003.

ZAHURAK, M. et al. Pre-processing agilent microarray data. *BMC Bioinformatics*, Springer Nature, v. 8, n. 1, p. 142, 2007. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186-/1471-2105-8-142>.

ZIMMERMANN, K.; LESER, U. Analysis of affymetrix exon arrays. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik, 2010.