

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

José Rafael Pilan

**Desenvolvimento de ferramentas de biologia de sistemas para
avaliação de dados transcriptômicos**

Botucatu, setembro de 2022

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

José Rafael Pilan

**Desenvolvimento de ferramentas de biologia de sistemas para
avaliação de dados de transcriptômicos**

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

Área de Concentração: Biotecnologia
Orientador: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, setembro de 2022

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

Pilan, José Rafael.

Desenvolvimento de ferramentas de biologia de sistemas para avaliação de dados de transcriptômicos / José Rafael Pilan. - Botucatu, 2022

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: José Luiz Rybarczyk Filho

Capes: 90400003

1. Biotecnologia. 2. Computação. 3. Proteínas. 4. Dados de expressão. 5. Biologia.

Palavras-chave: Biotecnologia; Computação; Dados de expressão; Proteínas; Redes.

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 pela orientação e auxílio;
- À toda minha família, principalmente meu pai José Carlos, minha mãe Maria Aparecida, meu irmão César e minha irmã Elaine pelo apoio e incentivo que foram fundamentais para a conclusão desse trabalho;
- Aos meus filhos Miguel e Marina por serem minha motivação diária em sempre buscar fazer o meu melhor;
- Aos meus amigos André Luiz Molan e Giordano Bruno Sanches Seco por toda a ajuda e motivação;
- À pós-graduação em Biotecnologia do Instituto de Biociências da Unesp de Botucatu e ao seu coordenador Prof. Dr. Valber de Albuquerque Pedrosa representando todos os docentes do programa;
- Aos demais colegas do departamento de Física e Biofísica e do Instituto de Biociências, professores e funcionários que de alguma forma tenham contribuído para a realização deste trabalho.

Resumo

A cada dia são desenvolvidos novas metodologias e equipamentos que proporcionam maior facilidade para a elucidação dos RNAs transcritos em situações específicas nos mais diferentes organismos. O que essas metodologias e equipamentos possuem em comum é o enorme volume de dados que podem gerar em cada execução. Devido a grande quantidade de tecnologias que estão surgindo para geração de dados de sequenciamento e expressão gênica precisamos de ferramentas que permitam o estudo e análise integrada dos diferentes dados advindos dos resultados desses experimentos. Nesse trabalho propomos a integração das funções do transcriptograma com uma ferramenta desenvolvida para visualização de dados de expressão sobreposto a uma rede biológica gerando como resultado final um pacote para execução no *layout* de *dashboard* das ferramentas citadas. O pacote foi desenvolvido utilizando a linguagem de programação R em conjunto com as linguagens C e C++ e disponibilizado para o ambiente R. Para validar a ferramenta utilizaremos dados de expressão gênica de amostras de pulmão obtidas das tecnologias de sequenciamento microarranjo e Single-Cell RNA. Como resultados a ferramenta *levi* (*Landscape Expression Visualization Interface*) está disponível para *download* em um dos principais repositórios *online*s de ferramentas de bioinformática, o Bioconductor. A integração das ferramentas do transcriptograma foram adicionadas ao *dashboard* e para validarmos a aplicação realizamos o *workflow* do transcriptograma em dados de microarranjo e RNA-seq de amostras de pessoas com e sem adenocarcinoma de pulmão. Os resultados indicam uma maior acurácia na utilização da ferramenta rTranscriptograma tanto para dados de microarranjo como para dados de RNA-seq.

Abstract

Every day new equipment are developed that provide greater facility for elucidating the RNAs transcribed in specific situations in the most different organisms. What these methodologies and equipment have in common is the large volume of data that they can generate in each execution. Due to the great amount of technologies that are emerging to generate sequencing and gene expression data, we need tools that allow the study and integrated analysis of the different data resulting from the results of these experiments. In this work we propose the integration of the functions of the transcriptogram with a tool to be developed for visualization of expression data superimposed on a biological network, resulting in a package for the execution of the mentioned tools in the final result. The package was developed using the R programming language together with C and C++ languages and made available to the R environment. To validate the tool we will use gene expression data from lung samples obtained from microarray and Single-Cell RNA sequencing Technologies. As a result the levi tool (Landscape Expression Visualization Interface) is available for download in one of the main online repositories of bioinformatics tools, the Bioconductor. The integration of transcriptogram tools has been added to the dashboard and to validate the application, we performed the transcriptogram workflow on microarray data and RNA-seq from samples from people with and without lung adenocarcinoma. The results indicate a greater accuracy in the use of the rTranscriptogram tool both from microarray and RNA-seq data.

Lista de Figuras

2.1	6 "V"s do <i>Big Data</i>	p. 4
2.2	Biologia de Sistemas	p. 5
2.3	Métodos de análises em Biologia de Sistemas	p. 7
2.4	Affymetrix GeneChip - Microarranjo de oligonucleotídeos.	p. 9
2.5	Roche NimbleGen - Microarranjo de oligonucleotídeos.	p. 10
2.6	Agilent - Microarranjo de oligonucleotídeos.	p. 11
2.7	Visão geral de um <i>pipeline</i> para RNA-seq	p. 12
2.8	Tecnologia de sequenciamento de Sanger.	p. 14
2.9	Estratégias de imobilização de <i>templates</i> utilizada pela Applied Biosystems. .	p. 15
2.10	Estratégia de amplificação por fase sólida usada pelos equipamentos da Illumina.	p. 16
2.11	Estratégia de fixação da polimerase no suporte utilizada pela Pacific Biosciences.	p. 17
2.12	Isolamento de <i>Single Cell</i> e preparação da biblioteca.	p. 19
2.13	Exemplo de rede direcionada e não-direcionada.	p. 23
2.14	Exemplos de tipos de grafos	p. 23
2.15	Principais modelos de redes	p. 26
4.1	<i>Workflow</i> com as etapas do projeto	p. 31
4.2	<i>Workflow</i> <i>levi</i>	p. 36
4.3	Transformação da rede em matriz de adjacência	p. 38
4.4	Método	p. 39
4.5	Proximidade dos vizinhos em relação ao elemento central na matriz de adjacência	p. 40

4.6	Representação de cinco possíveis distâncias do elemento central em relação a diagonal principal de uma matriz de adjacência.	p. 41
4.7	Avaliação da vizinhança do elemento central no modelo “cruz”.	p. 42
4.8	Gráfico de nível de custo energético por configurações possíveis de uma matriz de adjacência.	p. 43
4.9	Avaliação da vizinhança do elemento central no modelo “X”.	p. 44
4.10	Avaliação da vizinhança do elemento central no modelo “Anel”.	p. 45
4.11	Modularidade por Janela	p. 46
4.12	Resultado da Modularidade por Janela	p. 47
5.1	Rede do String para via de sinalização ErbB	p. 53
5.2	<i>Network</i> levi	p. 54
5.3	Objeto Ordering Data	p. 56
5.4	Objeto Modularity Data	p. 57
5.5	Tela principal de execução do <i>dashboard</i>	p. 59
5.6	Tela da opção <i>Cluster plot</i>	p. 60
5.7	Área de visualização do gráfico gerado a partir dos <i>clusters</i>	p. 60
5.8	Tela inicial da opção do menu <i>Gene expression Level</i>	p. 61
5.9	Aba <i>Biomart and Suavization</i> da opção do menu <i>Gene Expression Level</i> . . .	p. 63
5.10	Aba <i>Plot Settings</i>	p. 64
5.11	Gráficos de nível de expressão gênica - agrupamento	p. 64
5.12	Gráficos de nível de expressão gênica - unitário	p. 65
5.13	Aba <i>Targets</i> do menu <i>Functional Enrichment</i>	p. 65
5.14	Aba <i>Ontology and Statistic</i> do menu <i>Functional Enrichment</i>	p. 66
5.15	Aba <i>Comparison Results</i> do menu <i>Functional Enrichment</i>	p. 67
5.16	Aba <i>Graph</i> do menu <i>Functional Enrichment</i>	p. 67
5.17	Exemplo de visualização do levi dentro do <i>dashboard</i>	p. 68

5.18	Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo Cruz.	p. 69
5.19	Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo Anel.	p. 69
5.20	Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo X.	p. 70
5.21	Gráfico de Custo Energético em relação a quantidade de passos de Monte Carlo.	p. 70
5.22	Gráfico de modularidade por posição relativa das proteínas.	p. 72
5.23	Seleção de clusters para criação de sub-redes.	p. 72
5.24	Sub-redes criadas a partir da seleção dos módulos.	p. 74
5.25	Análise da atividade transcricional relativa da série GSE10072	p. 76
5.26	Análise da atividade transcricional relativa da série GSE10072	p. 76
5.27	Diagrama de Veen para a as comparações da série GSE10072	p. 77
5.28	Aplicação do levi para visualização da sobreposição do cluster AC em dados de expressão da série GSE10072.	p. 78
5.29	Análise da atividade transcricional relativa da série GSE10072	p. 79
5.30	Análise da atividade transcricional relativa da série GSE10072	p. 79
5.31	Diagrama de Veen para a as comparações da série GSE81089	p. 80
5.32	Aplicação do levi para visualização da sobreposição do cluster AC em dados de expressão da série GSE81089.	p. 81
5.33	Transcriptograma da comparação 1 (fumantes com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória. . .	p. 88
5.34	Transcriptograma da comparação 2 (ex-fumantes com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória.	p. 89
5.35	Transcriptograma da comparação 3 (indivíduos que nunca fumaram com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória.	p. 90

5.36	Transcriptograma da comparação 1 (fumantes com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória. .	p. 92
5.37	Transcriptograma da comparação 2 (ex-fumantes com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória.	p. 93
5.38	Transcriptograma da comparação 3 (indivíduos que nunca fumaram com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória.	p. 94

Lista de Tabelas

4.1	Detalhes dos dados de expressão utilizados na análise	p. 30
5.1	Número de <i>downloads</i> por ano e mês do pacote <i>levi</i> no Bioconductor.	p. 55
5.2	Posições relativas das proteínas para formação dos <i>clusters</i> selecionados.	p. 73
5.3	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer fumantes da série GSE10072.	p. 82
5.4	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer ex-fumantes da série GSE10072.	p. 83
5.5	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer que nunca fumaram da série GSE10072.	p. 84
5.6	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer fumantes da série GSE81089.	p. 85
5.7	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer ex-fumantes da série GSE81089.	p. 86
5.8	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer que nunca fumaram da série GSE81089.	p. 87

Sumário

Resumo	p. iv
Abstract	p. v
1 Introdução	p. 1
2 Revisão da Literatura	p. 3
2.1 Big Data	p. 3
2.2 Biologia de sistemas	p. 5
2.3 Tecnologias para obtenção e análise dos dados de expressão gênica	p. 8
2.3.1 Microarranjo	p. 8
2.4 Tecnologias de Microarranjo	p. 8
2.4.1 Affymetrix	p. 8
2.4.2 Roche	p. 9
2.4.3 Agilent	p. 10
2.4.4 RNA-seq	p. 11
2.4.5 Sanger (primeira-geração)	p. 12
2.4.6 Sequenciamento de nova geração	p. 14
2.4.7 Single Cell	p. 17
2.5 Normalização dos dados	p. 20
2.6 Dados transcriptômicos	p. 20
2.6.1 <i>Sequence Read Archive</i> (SRA)	p. 20
2.6.2 <i>Gene Expression Omnibus</i> (GEO)	p. 20
2.7 Redes	p. 21

2.7.1	<i>Search Tool for the Retrieval of Interacting Genes (STRING)</i>	p. 21
2.7.2	STITCH	p. 22
2.8	Redes Biológicas	p. 22
2.8.1	Conectividade	p. 24
2.8.2	Coeficiente de clusterização para redes não direcionadas	p. 24
2.8.3	Modelos de Redes	p. 24
2.8.4	Rede Escala Livre	p. 25
2.8.5	Rede Hierárquica	p. 26
2.8.6	Justificativa	p. 27
3	Objetivos	p. 29
3.1	Objetivos específicos	p. 29
4	Material e Métodos	p. 30
4.1	<i>Workflow</i>	p. 30
4.2	levi	p. 30
4.2.1	R	p. 32
4.2.2	C++	p. 33
4.2.3	CSS	p. 33
4.2.4	Equações	p. 33
4.3	<i>transcriptograma</i>	p. 37
4.4	Método de ordenamento de redes em uma dimensão	p. 37
4.4.1	Modelo Cruz	p. 40
4.4.2	Modelo X	p. 43
4.4.3	Modelo Anel	p. 44
4.5	Modularidade	p. 45
4.6	Análise de Expressão Gênica	p. 47

4.7	Enriquecimento Funcional	p. 48
4.8	Desenvolvimento da ferramenta <i>dashboard</i>	p. 48
4.8.1	Método transcriptograma	p. 48
5	Resultados e Discussão	p. 50
5.1	levi	p. 50
5.1.1	Implementação	p. 50
5.1.2	Instalação	p. 51
5.1.3	Carregando os arquivos	p. 52
5.1.4	Execução do levi	p. 52
5.1.5	Bioconductor	p. 55
5.2	transcriptograma	p. 55
5.3	Ferramenta de <i>dashboard</i>	p. 57
5.4	Aplicação do <i>workflow</i> em dados de <i>Homo sapiens</i>	p. 67
6	Conclusões	p. 96
	Referências Bibliográficas	p. 97

1 *Introdução*

A tecnologia da informação está cada vez mais presente em todas as áreas e aspectos de nossas vidas. Segundo (ALHARTHI, 2017), toda a atividade humana gera dados. No nosso dia a dia geramos dados nas interações que realizamos em redes sociais, com os equipamentos de GPS de celulares, ao acessar a redes *wi-fi* ou mesmo enviar mensagens em correios eletrônicos. Na área biológica, dados podem ser gerados de diferentes fontes como, por exemplo, os provenientes da interação entre proteínas ou os resultados obtidos através de tecnologias para o sequenciamento dos dados de expressão.

A integração e análise de dados computacionais obtidos de diferentes fontes é chamado de *big data* (SAGIROGLU; SINANC, 2013). Com o *big data* podemos realizar análises mais robustas incluindo dados que podem complementar resultados oriundos das mais diversas tecnologias.

Existem várias características do *big data* tal como os Vs que estão relacionados a atributos importantes que devemos levar em consideração para a correta análise dos diferentes conjuntos (variedade de dados) e a grande quantidade de dados existentes (volume). Para observarmos o nível de complexidade dos conjuntos, em uma análise de *big data* podemos exemplificar que a variedade dos dados podem ser estruturados, semi-estruturados ou não estruturados, um exemplo de dados não estruturados são as lojas virtuais que podem vender itens com diferentes características como tamanho, dimensão, volume, quantidade, cores, funcionalidades, entre outros. Em relação a característica do volume de dados podem ser medidos em terabytes, pentabytes e pode chegar até a *yottabytes* (RISTEVSKI; CHEN, 2018).

Para armazenar essa grande quantidade de dados de diferentes tipos existem repositórios disponíveis na *internet* como o IBGE com dados do senso brasileiro, o DATASUS com conjuntos de dados sobre o SUS e o Kaggle que possui dados desde filmes do cinema até desastres ambientais. Para dados obtidos através de diversas tecnologias ou experimentos biológicos podemos citar alguns exemplos como o *Gene Expression Omnibus*, *ArrayExpress* e o *The Cancer Genome Atlas*.

Para que possamos estudar a possibilidade de transferir as análises realizadas utilizando as técnicas de *big data* para a área biológica, podemos utilizar a biologia de sistemas. Segundo (SONIS; VILLA, 2019), a biologia de sistemas permite a integração entre áreas como a biologia, tecnologia e computação. As perguntas biológicas geram a necessidade de novas tecnologias, a aplicação das novas tecnologias tem como resultados novos dados e a computação, através de novos programas, realizam as análises desses dados.

Entre os grandes desafios da análise desses grandes volumes de dados estão as seleções dos conjuntos, integração e análises dos dados (KADADI et al., 2014). Por trabalhar com um grande volumes de dados que podem ter sido gerados de diferentes tecnologias, a etapa de análise de dados de *big data* apresenta uma grande dificuldade para os profissionais de ambas as áreas biológicas e/ou computacionais. Uma das principais etapas da análise de dados são as criações de visualizações através da filtragem e processamentos dos dados.

Apesar de tabelas possuírem geralmente um grau de detalhamento maior em relação a leitura dos dados, as imagens e gráficos facilitam a visualização e entendimento dos resultados por todos os usuários. Em 2017 Czauderna e Schreiber citam em seu trabalho a importância da visualização para apresentar e explorar grandes quantidades de dados biológicos.

Com a quantidade crescente de dados experimentais gerados e disponibilizados em repositórios na internet, faz-se necessário o acesso a métodos que permitam uma melhor visualização das potenciais relações entre estes registros.

2 *Revisão da Literatura*

2.1 **Big Data**

O crescente aumento da quantidade de conjuntos de dados gerados pelo surgimento de novas tecnologias, tanto de sequenciamento de nova geração quanto para análise de dados de expressão gênica, mais rápidas e a baixo custo permitiram o advento da era da ciência de *Big Data* (COSTA, 2014). O *Big Data* é a junção de um grande volume de dados, de diferentes tipos e que podem ser relacionados para gerar informação. Realizar esse tipo de análise, na maioria das vezes, não é algo trivial. Podem ser utilizados conceitos e técnicas de áreas relacionadas como inteligência artificial ou aprendizado de máquina para facilitar aos usuários a detecção de possíveis correlações nessa grande quantidade de dados para geração de informação.

O *Big Data* pode ser aplicado nas mais difentes áreas de interesse como no âmbito empresarial, instituições de ensino, dados médicos, finanças ou mesmo em dados de mídias sociais. Segundo (RISTEVSKI; CHEN, 2018) as características do *Big Data* podem ser definidas por 6 "V"s (figura 2.1):

- Volume: Pode ser medido em *terabytes*, *exabytes* ou mesmo *yottabytes*. É a quantidade de dados obtidos e que podem ser classificados e comparados para transformar dados em informação.
- Veracidade: Os dados podem ser avaliados em relação a sua qualidade, incerteza, relevância e confiabilidade.
- Velocidade: É o tempo de criação, processamento e análise dos dados.
- Variedade: Os dados podem estar estruturados, semi-estruturados ou não estruturados. A variedade leva em consideração a complexidade e heterogeneidade de todos os conjuntos de dados.
- Variabilidade: Os dados podem ser dinâmicos, estarem em evolução, serem sazonais. Todas essas variáveis são importantes para a análise correta dos dados.

- Valor: Procura mensurar a importância e a relevância dos dados.



Figura 2.1: Características do *Big Data*: Velocidade, Variedade, Volume, Variabilidade, Veracidade e Valor. Fonte: O Autor, 2022.

Entre as suas diversas utilizações é possível usar uma quantidade massiva de variáveis analisadas em conjunto para obtenção de modelos preditivos (KHOURY; IOANNIDIS, 2014). Esses modelos permitem prever os possíveis resultados levando em consideração as variáveis que permeiam o processo onde os dados estão inseridos. Esses modelos preditivos podem ser utilizados na área biológica, por exemplo, na fase de testes pré-clínicos da aplicação de drogas no combate a doenças permitindo analisar suas possíveis consequências.

Diversos estudos utilizaram o *Big Data* para auxiliar pesquisas médicas. Em 2019 Carbone e Montecucco realizaram um estudo sobre as oportunidades da utilização do *Big Data* para os desafios de doenças cardiovasculares. Brown e colaboradores utilizaram o *Big Data* para demonstrar que é possível melhorar o cronograma de projetos de descoberta de novas drogas além de otimizar as tomadas de decisão ((BROWN et al., 2018)). Sanches-Pinto e colaboradores realizaram um levantamento sobre aplicações, conceitos, desafios do *Big Data* em terapia intensiva (SANCHEZ-PINTO; LUO; CHURPEK, 2018).

2.2 Biologia de sistemas

No final do século XX e início do século XXI, o avanço em diversas áreas como nanotecnologia, robótica, processamento de dados e óptica permitiram a evolução das ferramentas de análise moleculares a um nível mais amplo, possibilitando assim novos estudos dos organismos. Essa evolução gerou as chamadas "Ciências Ômicas" que buscam analisar ou mesmo comparar a estrutura e o comportamento de milhares de componentes biológicos em diferentes condições fisiológicas ou em diferentes espécies. Entre essas áreas podemos ter o estudo da sequência completa de DNA de um organismo (genoma), o conjunto de transcritos de RNA (transcriptoma), o estudo do conjunto de proteínas (proteoma), entre outros. (MOREIRA, 2015)

Os estudos ômicos geram uma vasta quantidade de dados e informações relacionadas ao comportamento celular em diversas condições fisiológicas. Para elucidar os elos das lacunas ausentes entre moléculas e fisiologia foi desenvolvida a biologia de sistemas. A biologia de sistemas busca integrar todos esses dados e informações mapeando e modelando em termos quantitativos, o comportamento celular ou fisiológico de um organismo (figura 2.2) (BRUGGEMAN; WESTERHOFF, 2007).

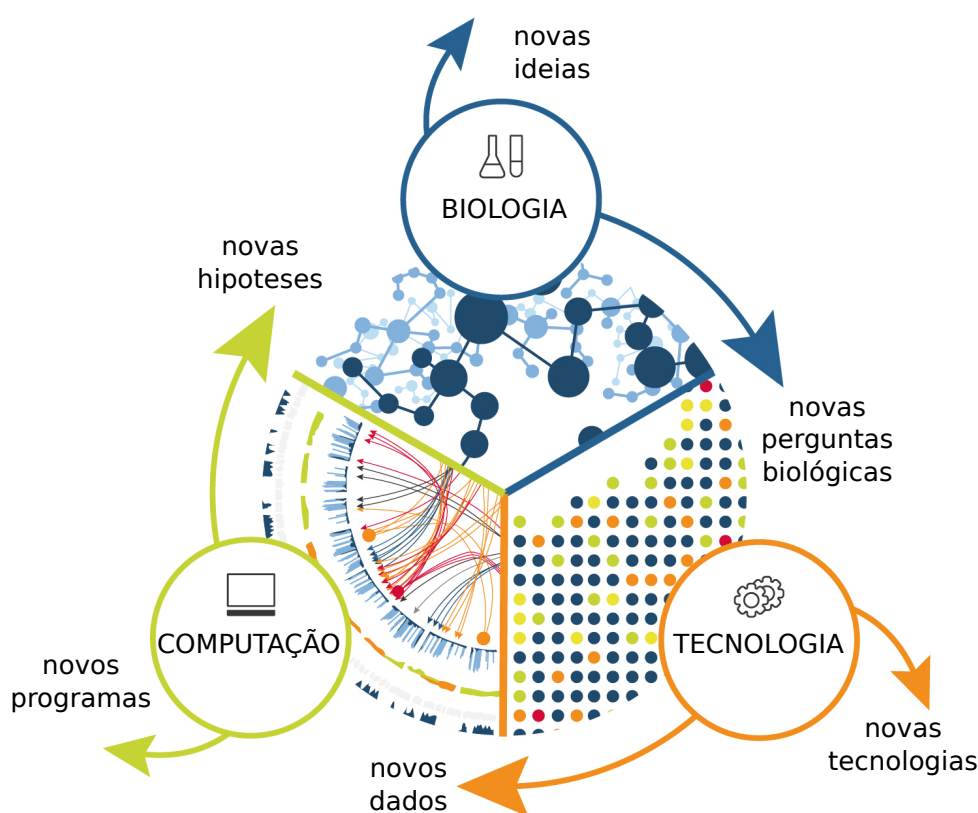


Figura 2.2: Biologia de Sistemas. O presente esquema integra áreas como Biologia, Tecnologia e Computação para desenvolvimento de novos insumos buscando responder perguntas biológicas. Fonte: Adaptado de (SONIS; VILLA, 2019)

Buscando atingir esse objetivo, uma das principais metodologias computacionais que pode ser utilizada para integrar todos esses dados são conceitos existentes na Teoria dos Grafos, um ramo da matemática que lida com os fundamentos das chamadas redes de integração. Uma rede de integração é uma estrutura complexa, que pode ser formada por um elemento chamado nós (ou vértices), que são unidos por conectores e auxiliam a descrever uma variedade de sistemas biológicos, de tecnologia ou relações sociais. Por exemplo, a internet é uma complexa rede de computadores e roteadores (nós), unidos por meios físicos (conectores) (ALBERT; BARABÁSI, 2002).

Segundo (SHAHZAD; LOOR, 2012) existem dois métodos de análises em biologia de sistemas que levam em consideração os diferentes tipos de redes biológicas: a biologia de sistemas "de cima para baixo" (que realiza o estudo do mais complexo ao menos complexo ou "de cima-para-baixo") e a biologia de sistemas "de baixo para cima" (indo do menos complexo ao mais complexo ou "de baixo-para-cima"). A análise "de cima para baixo" utiliza os dados experimentais de larga escala obtidos pelas ciências ômicas, utilizando esses dados é possível realizar análises e integração destes dados para a geração de novas hipóteses. Estas hipóteses podem ser testadas através da verificação experimental do modelo (Fig 2.3 A). A análise "de baixo para cima" inicia-se utilizando de processos já caracterizados experimentalmente (reações enzimáticas, interações proteicas, etc), a segunda etapa é integrar esses processos computacionalmente em diferentes contextos biológicos. As hipóteses criadas através dessas análises são verificadas experimentalmente e podem fornecer uma visão mais ampla sobre o funcionamento dos sistemas biológicos (Fig 2.3 B).

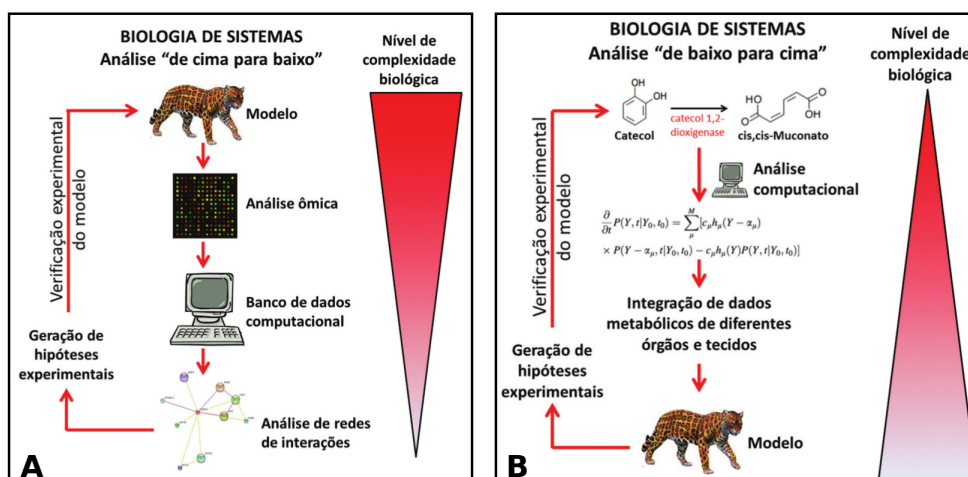


Figura 2.3: Métodos de análises em Biologia de Sistemas. (A) Análise "de baixo-para-cima". Este método utiliza os dados experimentais de larga escala obtidos através das ciências ômicas para elaboração das hipóteses. (B) Análise "de baixo para cima". Este método busca integrar de forma computacional os processos já caracterizados experimentalmente para a geração das suas hipóteses. Fonte: Adaptado de (MOREIRA, 2015)

Para avaliar se os resultados obtidos em modelos animais podem ser aplicados para o *Homo sapiens* podemos utilizar a biologia de sistemas. A biologia de sistemas provê meios para identificar a que ponto os mecanismos estudados em modelos *in vivo* são similares. Ela é uma ciência multidisciplinar que utiliza abordagens matemáticas com percepções biológicas e médicas com o intuito de predizer eventos interconectados para análise da rede de interações utilizando diferentes níveis do sistema (GONZALEZ-ANGULO; HENNESSY; MILLS, 2010).

A biologia de sistemas oferece os meios para os limites de tradução de modelos animais em diferentes ensaios clínicos às avaliações toxicológicas para a biologia celular. Pois, podemos integrar os dados contidos em interações entre proteínas, rotas metabólicas, redes, ontologias, expressão gênica e pequenas moléculas para construir modelos matemáticos preditivos desse sistema e assim compreender essas informações em diferentes níveis dos sistemas biológicos individuais (IDEKER; GALITSKI; HOOD, 2001).

Com a utilização do *Big Data* podemos analisar dados entre interações de RNA longo não-codificante (*Long non-coding RNA* - lncRNA), microRNA (*miRNA*) e proteínas de bancos de dados *online*. Os lncRNA são uma classe de RNAs não codificadores de proteínas que possuem mais de 200 nucleotídeos, estão envolvidos em uma ampla variedade de processos biológicos como moléculas reguladoras (SZCZEŚNIAK; MAKALOWSKA, 2016). Os microRNAs (miRNAs) são classes caracterizadas de ncRNAs e são pequenos reguladores de moléculas de RNA a partir de precursores maiores através de uma via altamente coordenada (JALALI et al., 2013).

2.3 Tecnologias para obtenção e análise dos dados de expressão gênica

2.3.1 Microarranjo

A tecnologia de microarranjo foi um dos grandes benefícios obtidos durante pesquisas realizadas no decorrer do Projeto do Genoma Humano (TREVINO; FALCIANI; BARRERA-SALDANA, 2007). Com a utilização dessa tecnologia, tornou-se possível mensurar, ao mesmo tempo com rapidez e eficiência, a expressão de vários genes de um determinado organismo. O princípio da técnica de microarranjo *in situ* consiste na utilização de uma superfície sólida que possui diversas sequências de nucleotídeos com tamanhos que variam dependendo do chip utilizado (entre 20 a 100 pares de bases por sonda) fixadas em sua base, essas sequências podem ser de dois tipos: cDNA ou oligonucleotídeos (MILLER; TANG, 2009). Os oligonucleotídeos ou o cDNA são usados como sonda que quando híbrida com uma molécula alvo marcada com fluoróforo permite a detecção de DNA complementar ou RNA. Em pouco tempo, o microarranjo se tornou uma das ferramentas mais utilizadas em meta análises de dados genômicos (TSENG; GHOSH; FEINGOLD, 2012).

2.4 Tecnologias de Microarranjo

2.4.1 Affymetrix

Na tecnologia de microarranjo da Affymetrix os oligonucleotídeos são sintetizados diretamente na superfície do microarranjo. É uma superfície de quartzo com proporções de 1,2cm². O processo de sintetização *in situ* cria sondas normalmente curtas, entre 20 a 25 pares de base, por isso é utilizado várias sondas para se obter melhores valores estatísticos de sensibilidade, especificidade e precisão. (DALMA-WEISZHAUSZ et al., 2006).

O processo de adição das bases A, C, T ou G na cadeia de oligonucleotídeos é realizada utilizando uma máscara litográfica e luz UV. A luz UV remove a proteção química existente no microarranjo, permitindo assim a síntese de apenas um nucleotídeo por vez. Padrões diferentes de utilização de máscaras em cada ciclo liberam sondas diferentes para a adição dos oligonucleotídeos em determinadas cadeias. A execução e repetição desse processo gera o resultado final com sondas de até 25 pb (figura 2.4).

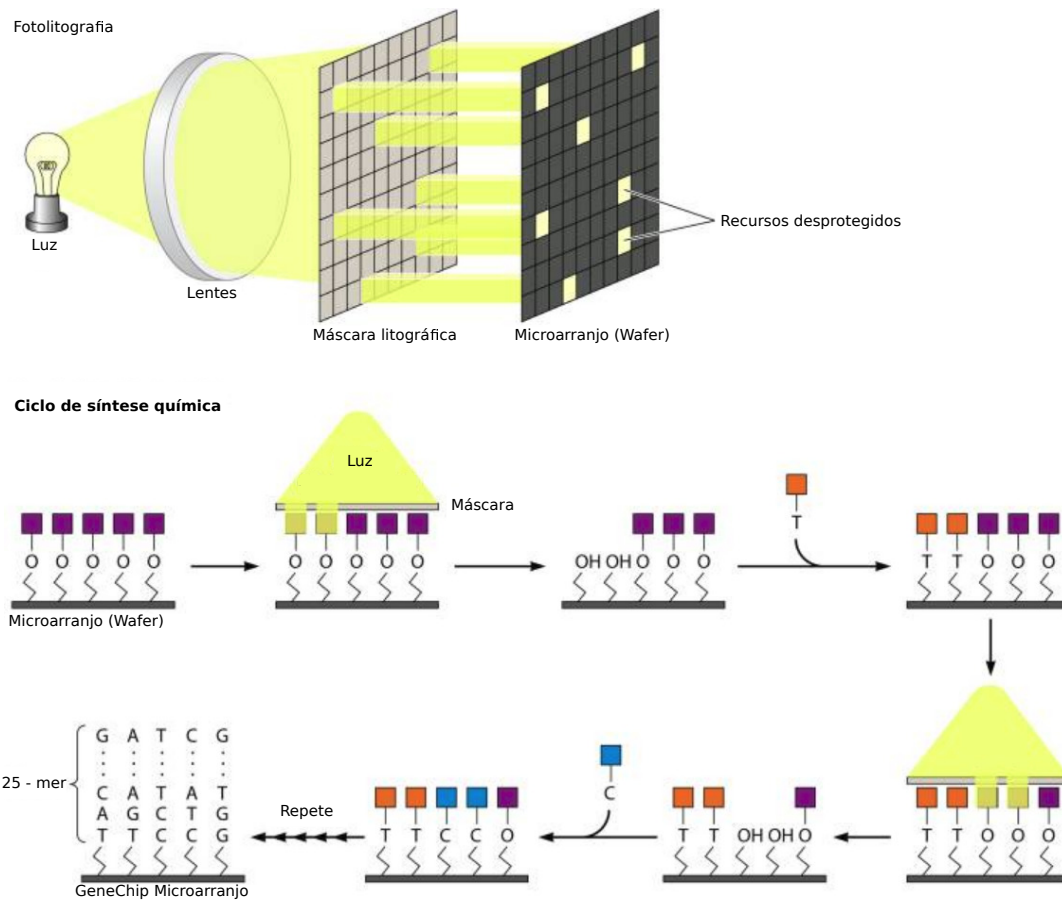


Figura 2.4: O esquema da Affymetrix utiliza uma fotolitografia a luz UV que emite uma luz incidindo na máscara litográfica que realiza o filtro para transmitir ou bloquear a luz no microarranjo que está quimicamente protegido. A ordem da síntese da sequência na superfície do microarranjo é determinada com a aplicação sequencial de máscaras litográficas específicas. No Ciclo de síntese química a luz UV remove os grupos de proteção (quadrados) da superfície da matriz, permitindo a adição de um único nucleotídeo protegido à medida que ela é lavada sobre o microarranjo. A aplicação de sequências de remoção de proteção da luz, alterações nos padrões de filtragem das máscaras e adições de nucleotídeo único formam o microarranjo com sondas específicas de tamanho de 25 pb. Adaptado de (MILLER; TANG, 2009).

2.4.2 Roche

A tecnologia de microarranjo da Roche consegue criar sondas de oligonucleotídeos maiores que a da Affymetrix com tamanhos entre 50 a 100 pb. Outra diferença entre os microarranjos da Affymetrix e da Roche é a utilização das máscaras. Enquanto a Affymetrix utiliza máscaras físicas a Roche utiliza máscaras virtuais. Essas máscaras são criadas através de um Dispositivo Digital de Microespelhos (*Digital Micromirror Device* - DMD) que é responsável por canalizar a luz na superfície do microarranjo removendo a proteção para que o possível nucleotídeo seja adicionado (figura 2.5). (HUANG; HUANG; WANG, 2010)

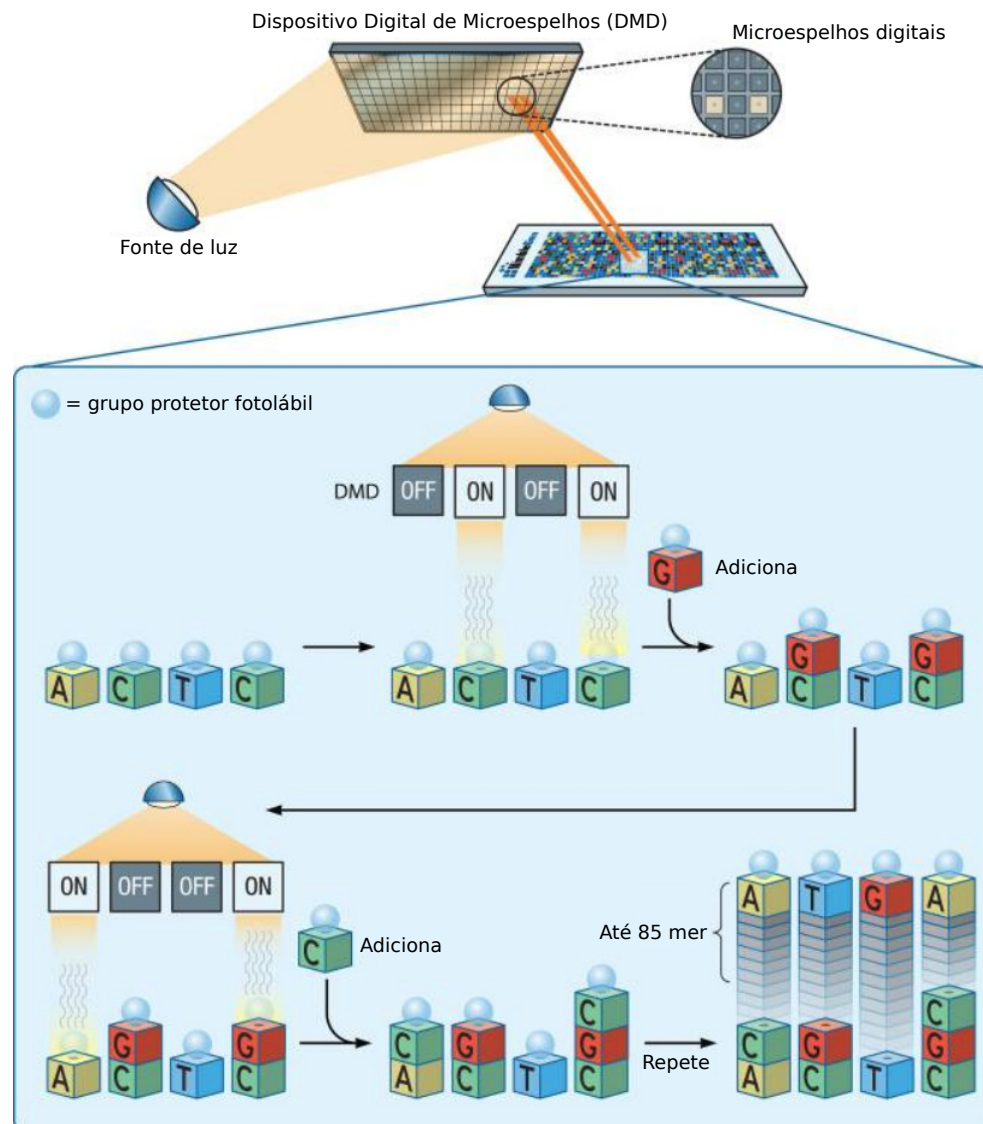


Figura 2.5: O esquema da Roche utiliza um sintetizador de matriz que não precisa de máscara física. Com a utilização de um DMD - *Digital Micromirror Device* (Dispositivo Digital de Microespelhos) são criadas máscaras virtuais. O DMD direciona a luz UV permitindo que seja adicionado apenas os ácidos nucleicos específicos durante a etapa de síntese. A luz UV remove o grupo protetor fotolábil da superfície do microarranjo, permitindo a adição de um único nucleotídeo protegido à cadeia oligonucleotídica. Essa tecnologia permite que sejam criados de 60 a 100 bp de comprimento no microarranjo. Adaptado de (MILLER; TANG, 2009).

2.4.3 Agilent

As sondas sintetizadas pela tecnologia *in situ* da Agilent utiliza impressão com jatos de tinta. Essa tecnologia permite a hibridização com múltiplas cores, cada cor correspondente aos nucleotídeos A, C, T e G. Diferente dos suportes da Affymetrix e da Roche, os suportes

da Agilent são feitos de vidro e para a construção do equipamento são utilizadas cabeças de impressão a jato de tinta comerciais. Para a impressão são usados monômeros de fosforamidita em carbonato de propileno (figura 2.6). (HUGHES et al., 2001)

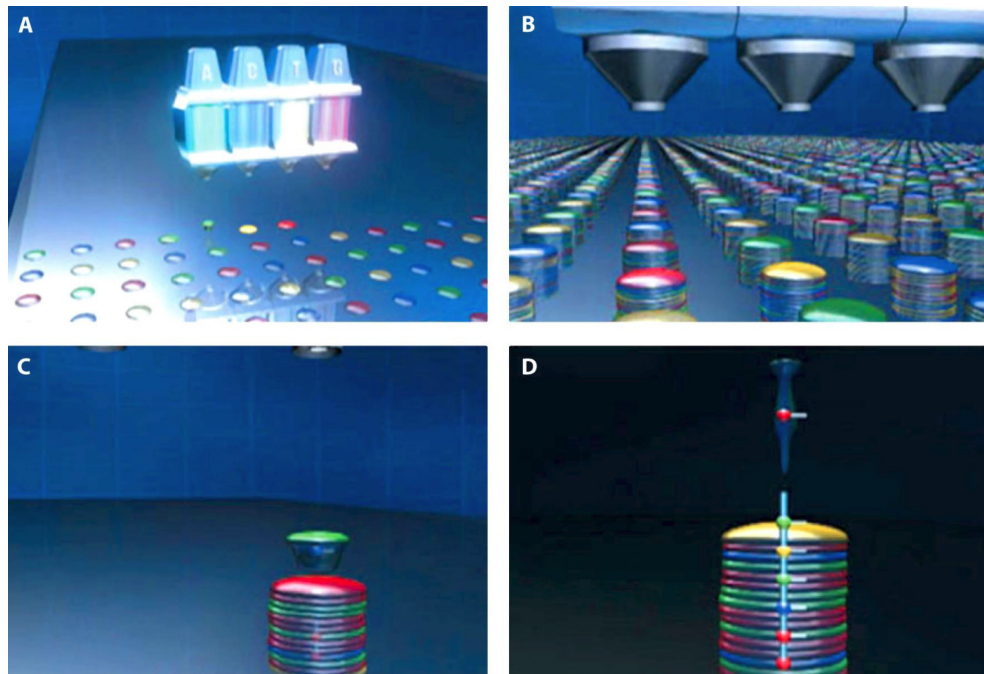


Figura 2.6: No esquema da Agilent temos: (A) Usando uma tecnologia de impressão de tinta não é necessário contato direto, assim é gerada a primeira camada de nucleotídeos na superfície do microarranjo. (B) Esse procedimento se repete aumentando o comprimento das sondas oligonucleotídicas específicas. (C) Visão aproximada da cadeia oligonucleotídica sendo aumentada base a base. (D) A sonda sintetizada *in situ* pela Agilent possui comprimento de 60 pb. Adaptado de (MILLER; TANG, 2009).

2.4.4 RNA-seq

O RNA-seq é uma tecnologia que permite gerar o perfil do transcriptoma. O perfil do transcriptoma leva em consideração o conjunto e a quantia de transcritos de uma célula em uma determinada condição fisiológica ou em um estágio de desenvolvimento específico. Usando tecnologias de sequenciamento o RNA-seq permite mensurar de forma mais precisa o nível dos transcritos e suas isoformas. Entender o perfil do transcriptoma é fundamental para elucidar os elementos do genoma, compreender os constituintes moleculares das células e tecidos e assim permitir o estudo e desenvolvimento de doenças (WANG; GERSTEIN; SNYDER, 2009). O RNA-seq possui um menor ruído de *background* do que a tecnologias de microarrays e também possui uma faixa dinâmica maior para ser realizada a detecção (HRDLICKOVA; TOLOUE; TIAN, 2017).

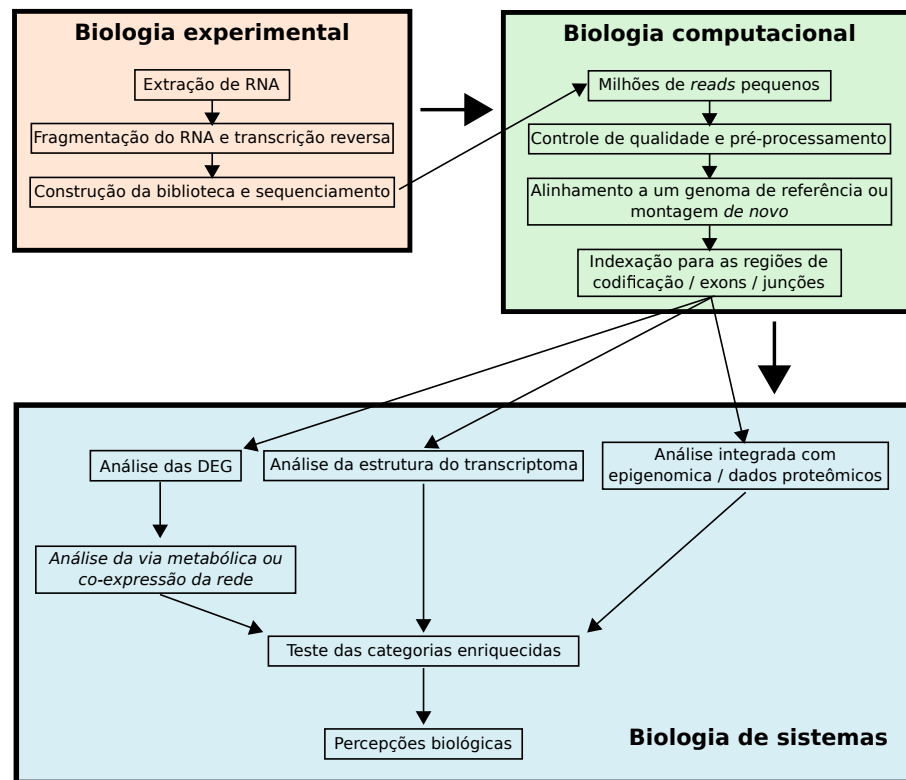


Figura 2.7: Esquema geral de um *pipeline* para RNA-seq. É dividido em três áreas principais: biologia experimental, biologia computacional e biologia de sistemas. O *pipeline* é iniciado com a extração de DNA na fase da biologia experimental, que é responsável pela preparação da amostra. Na fase de biologia computacional é realizado o controle de qualidade, alinhamento/montagem e indexação das regiões. Por último, temos a biologia de sistema com três etapas iniciais: as análises de genes diferencialmente expressos (DEG), a análise de estrutura do transcriptoma e as análises integradas. Todas elas levam à percepções biológicas. Adaptado de (HAN et al., 2015).

O RNA-seq realiza o seu processo baseado nas tecnologias de sequenciamento. Para cada geração de sequenciamento existem tecnologias e seus respectivos métodos utilizados. Entre as principais tecnologias temos: Sanger (primeira-geração), Illumina (segunda-geração) e Pacific Biosciences (terceira-geração).

2.4.5 Sanger (primeira-geração)

A tecnologia de sequenciamento criada por Sanger e colaboradores (SANGER; NICKLEN; COULSON, 1977), conhecida também como de “primeira-geração”, utiliza dideoxinucleotídeos como bloqueadores de síntese da cadeia de DNA (METZKER, 2010) (figura 2.8). Esse é um método de sequenciamento, onde ocorre a síntese de uma fita complementar a sequência de DNA de interesse, para isso são necessários DNA polimerase, o *primer* que servirá como ini-

ciador os deoxinucleotídeos (DNTPs) para a síntese e os dideoxinucleotídeos (ddNTPs) como finalizadores inibidores. Os dideoxinucleotídeos (ddNTPs) são nucleotídeos modificados que não possuem um grupo hidroxila na posição 3' do carbono do anel de sacarose bloqueando assim a adição de novos nucleotídeos. Durante a síntese, várias cópias da fita complementar são sintetizadas com a adição de DNTPs. Porém, quando os ddNTPs são incorporados, ocorre a interrupção da inclusão de novos nucleotídeos. Essa interrupção pode ocorrer em qualquer etapa da síntese, gerando fragmentos de diferentes tamanhos de dNTPs. Após essa etapa, os fragmentos de DNA são separados utilizando um gel de poliacrilamida. A sequência é obtida através da visualização das bandas que indicará a sequência de nucleotídeos da fita complementar e, por complementaridade de bases. Atualmente o método passou a ser automatizado, com a utilização de computadores para leitura do gel e processamento das sequências e a utilização de fluorocromos para os nucleotídeos (HEATHER; CHAIN, 2016).

A utilização do método de Sanger permitiu o surgimento de novos tipos de análises onde o sequenciamento de vários organismos simultaneamente está sendo utilizado e possibilitando estudos aprofundados que o isolamento de materiais genéticos de um único organismo não propiciava. Utilizando o sequenciamento de Sanger foi possível, por exemplo, sequenciar o genoma humano (METZKER, 2010).

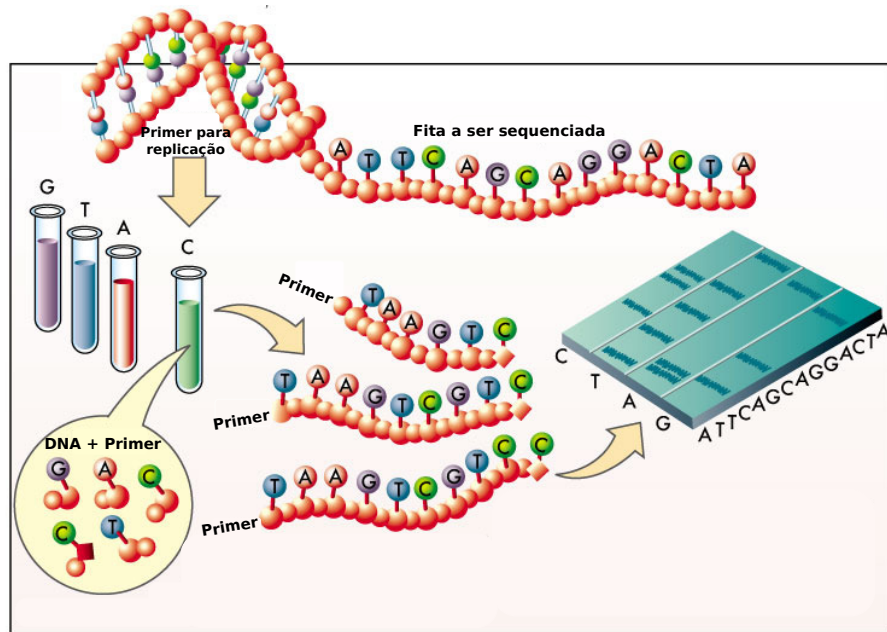


Figura 2.8: Ilustração do funcionamento da tecnologia de sequenciamento de Sanger. À temperatura de 95°C a fita dupla sofre uma desnaturação separando-se em duas fitas simples. Cada uma dessas fitas simples possui *primer*, nucleotídeos e DNA polimerase. Aproximadamente 50°C ocorre o anelamento do *primer* e com 72°C a DNA polimerase entra em ação realizando a extensão da fita complementar. Quando os dideoxynucleotídeos (ddNTPs) são incorporados à fita de DNA ocorre o bloqueio da adição de novos nucleotídeos, gerando assim fragmentos de tamanhos diferentes. Ao colocar no gel essas fitas migram por causa da sua massa e carga. No sequenciamento automatizado os ddNTPs (dideoxynucleotídeos) são utilizados utilizando um sistema de laser e detector sendo possível identificar esses nucleotídeos, que são fluorescentes, no capilar. Adaptado de (WINNICK, 2004).

2.4.6 Sequenciamento de nova geração

Os métodos de sequenciamento que surgiram após o método de Sanger e que utilizam estratégias como preparação de fitas molde, sequenciamento e imagem, alinhamento de genoma e métodos de montagem são chamados de sequenciamentos de nova geração *next-generation sequencing* (NGS). As empresas de equipamentos de NGS utilizam tecnologias diferentes para a criação dos *templates* (modelos), sequenciamento e captura de imagem. Essas tecnologias diferentes e seus métodos estão diretamente relacionadas à qualidade e ao custo dos dados gerados. (METZKER, 2010).

Entre as principais fabricantes de equipamento de sequenciamento temos: Applied Biosystems, Illumina, Pacific Biosciences, Ion Torrent, Oxford Nanopore e a SOLiD. Para exemplificar essas tecnologias veremos o funcionamento da metodologia aplicada pelas empresas Applied Biosystems, Illumina, Pacific Biosciences.

2.4.6.1 Applied Biosystems

A Applied Biosystems utiliza o método de preparação do *template* onde são fixados os *primers*, de forma individual e são distribuídos sobre um suporte sólido, esses suportes em formato de esferas são fixadas em uma lâmina de vidro (figura 2.9). Na tecnologia da Applied Biosystems, os tubos possuem uma mistura de *primers*, *templates*, dNTPs e polimerase. Ocorre uma reação de amplificação por PCR, que é quebrada por emulsão ocorrendo assim a separação do *template*.

Applied Biosystems

Emulsão por PCR

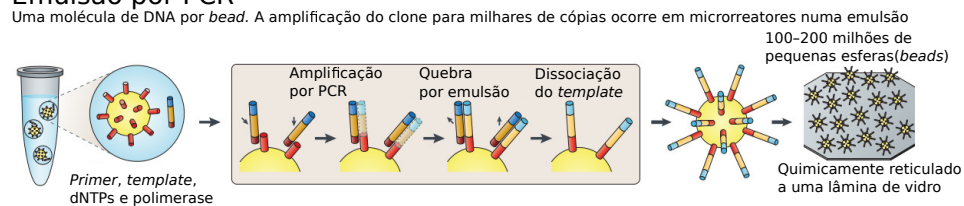


Figura 2.9: Estratégias de imobilização de *templates* utilizada pela Applied Biosystems. Primeiro ocorre a adição e ligação dos adaptadores em seguida são selecionados os fragmentos que possuem adaptadores. Em uma segunda etapa, acontece a ligação da fita simples de DNA às esferas e em seguida a amplificação do DNA por PCR. Por último, temos a eliminação das gotículas. Com a presença de reagentes (sulfurilase) é gerada luz que é captada pelo equipamento. Adaptado de (METZKER, 2010).

2.4.6.2 Illumina

A Illumina utiliza um método de fixação de uma única molécula de DNA por *cluster* (figura 2.10). Nessa técnica, são criados agrupamentos utilizando fragmentos ou *mate-paired templates* em uma lâmina de vidro. Cada agrupamento possui *primers* que durante a fase de amplificação criarão pontes entre os segmentos *forward* e *reverse* da fita de DNA.

Illumina

Amplificação por fase sólida

Uma molécula de DNA por *cluster*.

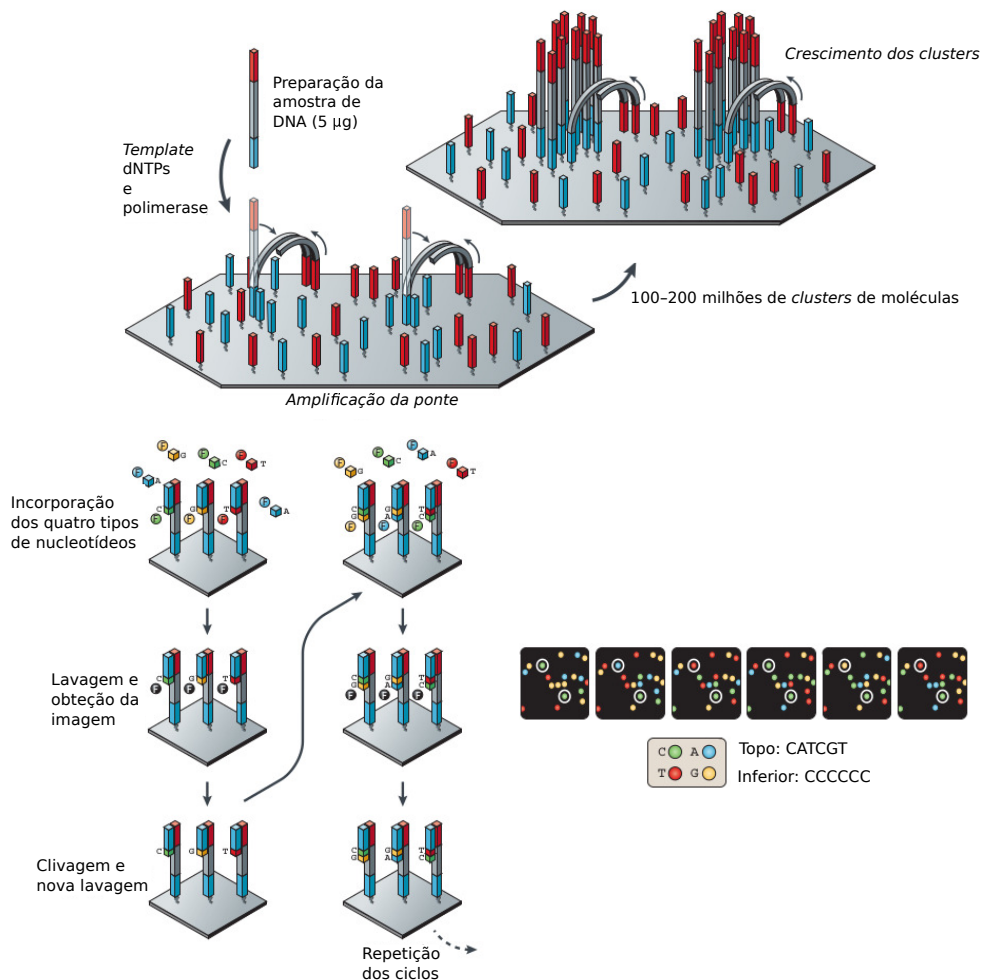


Figura 2.10: Estratégia de amplificação por fase sólida usada pelos equipamentos da Illumina. Em cada *cluster*, há a ligação entre os segmentos da fita de DNA *forward* e *reverse*. Com isso os nucleotídeos são incorporados para a criação da fita de DNA complementar. Ocorre uma lavagem para retirar os nucleotídeos não incorporados, é obtida a imagem pelo equipamento, é realizada a clivagem para retirar os fluorocromos (responsável pela emissão de luz). Esse ciclo é repetido até a obtenção de toda a sequência de nucleotídeos da fita complementar. Adaptado de (METZKER, 2010).

2.4.6.3 Pacific Biosciences

O método utilizado pela Pacific Biosciences realiza a detecção dos fragmentos incorporados em tempo real e é considerado de terceira geração. Nesse método, a polimerase é fixada à superfície sólida. A detecção dos nucleotídeos incorporados ocorre em tempo real com isso é possível conseguir tamanhos de *reads* maiores do que as outras metodologias (figura 2.11)(LIU et al., 2012; METZKER, 2010).

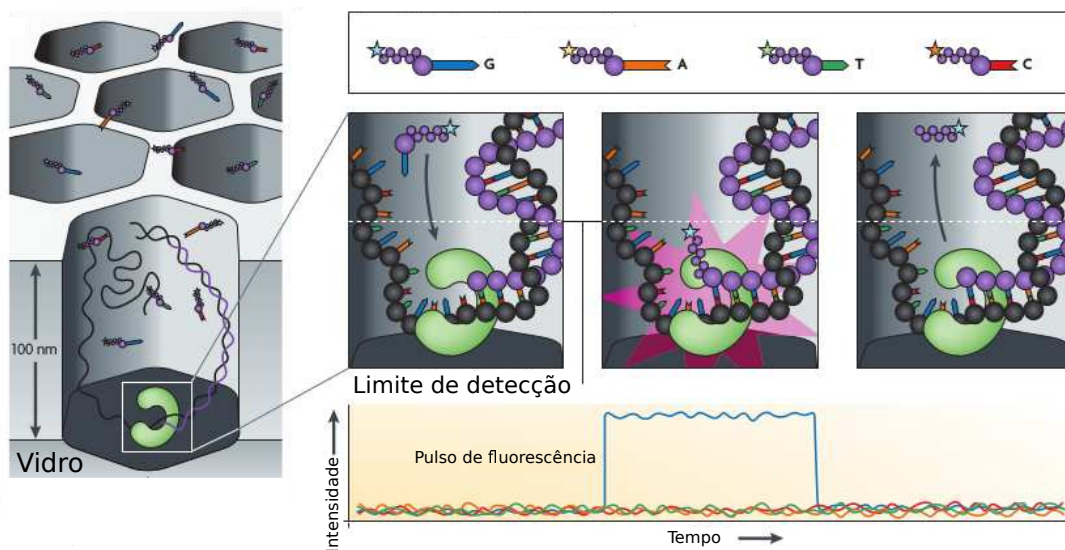


Figura 2.11: Estratégia de fixação da polimerase no suporte utilizada pela Pacific Biosciences. A polimerase é fixada em uma superfície sólida. O equipamento possui um limite de detecção para captura da imagem, com isso consegue verificar a intensidade em relação ao tempo do pulso de fluorescência quando um nucleotídeo é incorporado. Após a captura da imagem, ocorre a clivagem do fluorocromo. Adaptado de (METZKER, 2010).

2.4.7 Single Cell

Os resultados em uma análise de RNA-seq são obtidos utilizando os cálculos dos valores de perfis de expressão das médias de um conjunto de células. O viés da utilização de RNA-seq é a impossibilidade de realizar uma análise minuciosa nos diversos tipos e subtipos de células usando análises baseadas em população. Desse modo, a técnica de *single cell* permite entender a expressão gênica no nível de célula única o que é importante para estudos e análises isoladas dessas células. (HRDLICKOVA; TOLOUE; TIAN, 2017) (LUN; MCCARTHY; MARIONI, 2016)

O primeiro método de sequenciamento de RNA *single cell* (scRNA-seq) foi desenvolvido em 2009 utilizando a premissa que células individuais são únicas e a base para a construção dos

organismos. A execução de um sequenciamento em massa de RNA pode mascarar a singularidade de cada célula além de ocultar possíveis mudanças nessas células (TANG et al., 2009).

Para cada célula, o mRNA é isolado e transcrito reversamente em cDNA para execução do sequenciamento utilizando as tecnologias de nova geração (NGS - *Next Generation Sequencing*). Essa tarefa pode ser realizada usando plataformas microfluídicas como o Fluidigm C1, protocolos baseados em placas de microtitulação como Smart-seq2 ou tecnologias baseadas em gotículas como inDrop. A quantificação da expressão de cada célula é obtida através do mapeamento das suas leituras. Também podem ser utilizados identificadores moleculares únicos (UMIs) para medir o número de moléculas de transcrição para cada gene. Os dados de contagem são analisados e fornecem informações biológicas para detectar genes altamente variáveis (HVGs) que demonstram a heterogeneidade entre células em uma população, é possível encontrar correlações entre genes e fenótipos celulares assim como identificar novas subpopulações através da identificação de agrupamentos (figura 2.12) (LUN; MCCARTHY; MARIONI, 2016).

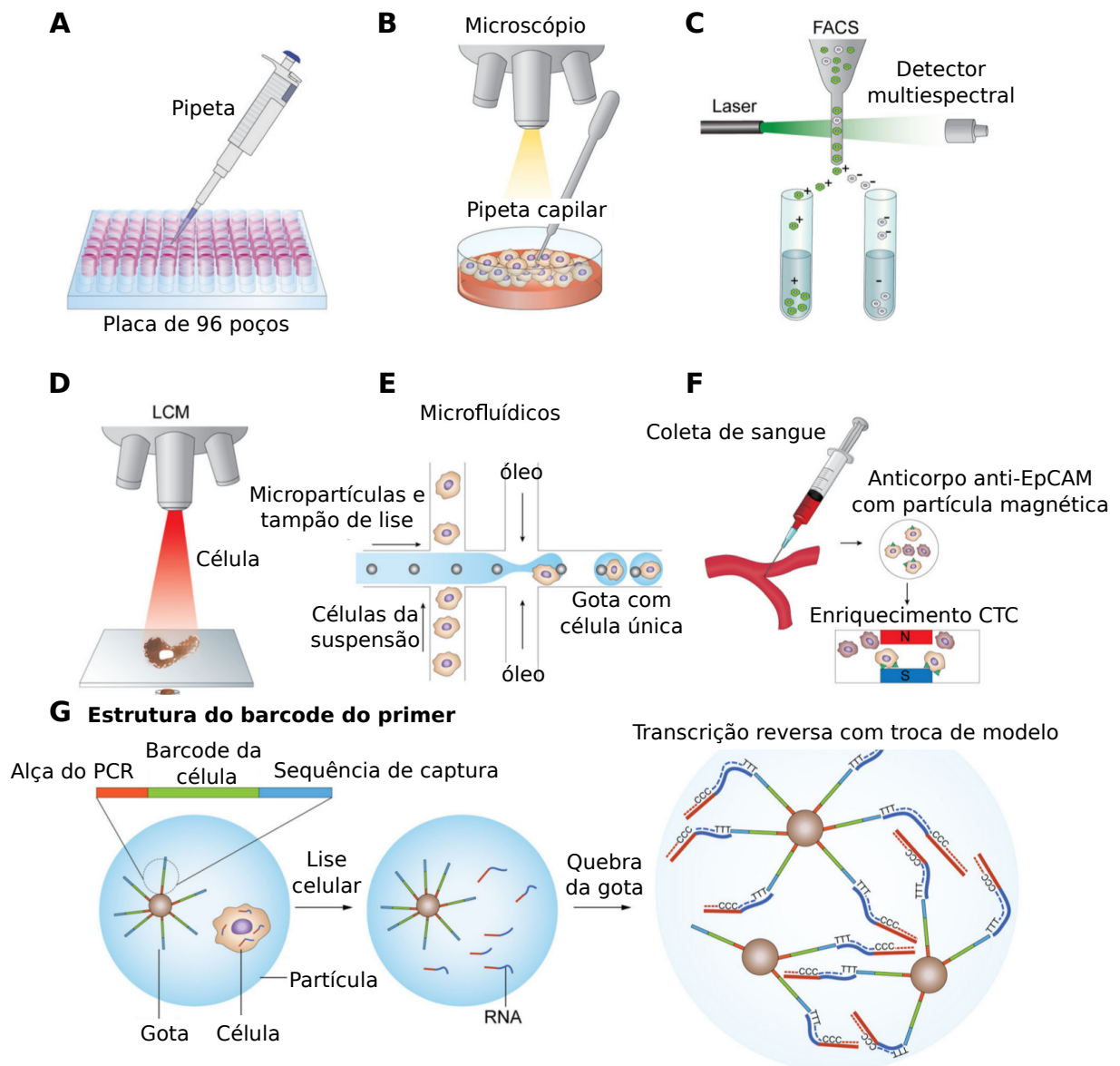


Figura 2.12: Isolamento de *Single Cell* e preparação da biblioteca. (A) O método de diluição limitante isola células individuais, alavancando a distribuição estatística de células diluídas. (B) A micromanipulação envolve a coleta de células únicas usando pipetas capilares guiadas por microscópio. (C) FACS (*Fluorescence Activated Cell Sorting*) isola as células altamente purificadas através da marcação com proteínas fluorescentes. (D) A microdissecação de captura a laser (*Laser capture microscopy* - LCM) utiliza um sistema de laser auxiliado por um sistema de computador para isolar células de amostras sólidas. (E) A tecnologia microfluídica para isolamento de célula única requer volumes do tamanho de nanolitros. (F) O sistema CellSearch enumera CTCs (*Circulating tumor cells*) de amostras de sangue de pacientes usando um ímã conjugado com anticorpos de ligação à CTC. (G) Um exemplo esquemático de geração de biblioteca baseada em gotículas. Bibliotecas para scRNA-seq geralmente são geradas via lise celular, transcrição reversa no cDNA da primeira fita usando barcodes exclusivos, síntese da segunda fita e amplificação de cDNA. Adaptado de (HWANG; LEE; BANG, 2018).

2.5 Normalização dos dados

Os dados de microarranjo precisam ser normalizados com a finalidade de diminuir as variações existentes entre os dados, essas variações ocorrem devido a diferença na eficiência de marcação dos fluoróforos verde e vermelho no processo de detecção no escaneamento. Com a normalização é possível realizar comparações apropriadas entre os dados (DUDOIT et al., 2002).

Para a os dados obtidos por meio de técnicas de RNA-seq podemos utilizar dois tipos de normalizações: RNA-seq Bulk e RNA-seq single cell. O protocolo RNA-seq Bulk é utilizado para o RNA-seq tradicional enquanto o RNA-seq single cell são utilizados para amostras obtidas pela técnica de single Cell. (LUN; MCCARTHY; MARIONI, 2016).

Os dados de microarranjo podem ser normalizados usando um dos dos 3 métodos/protocolos:

- *Microarray Suite 5 (MAS5)*;
- *Robust Multi-array Average (rma)*;
- *GC Robust Multi-array Average (gcrma)*;

2.6 Dados transcriptômicos

2.6.1 *Sequence Read Archive (SRA)*

O *Sequence Read Archive (SRA)* é um banco de dados público e internacional para armazenamento de dados de sequências obtidas de tecnologias de NGS (*Next Generation Sequence*). É administrado pelo *National Center for Biotechnology Information (NCBI)*, o Instituto Europeu de Bioinformática (EBI) e o *DNA Data Bank of Japan (DDBJ)*, tendo como missão preservar os dados de sequenciamento de domínio público e fornecer acesso livre, irrestrito e permanente a esses dados (LEINONEN; SUGAWARA; SHUMWAY, 2010). No SRA é possível obter dados de RNA-Seq.

2.6.2 *Gene Expression Omnibus (GEO)*

O GEO (*Gene Expression Omnibus*) possui dados de RNA-Seq e microarranjo. O GEO é um projeto que iniciou com o aumento exponencial de dados de expressão gênica e a necessi-

dade de um repositório público para esses dados (EDGAR; DOMRACHEV; LASH, 2002).

2.7 Redes

2.7.1 *Search Tool for the Retrieval of Interacting Genes (STRING)*

Para a criação das redes de integração utilizaremos dados obtidos no banco de dados STRING. O STRING é um banco de dados de interações de proteínas-proteínas que atualmente possui 5090 organismos (o dobro da versão anterior 2031 organismos), 24.6 milhões proteínas e mais de 2 bilhões de interações cadastradas (STRING, 2017). Segundo Szklarczyk e colaboradores o STRING objetiva fornecer uma avaliação crítica e integração de interações proteína-proteína, incluindo associações diretas (físicas) e indiretas (funcionais) ((SZKLARCZYK et al., 2014)). Na sua versão mais atual (STRING V11), possui melhorias que incluem a possibilidade de carregar como *input* um conjunto de dados de genoma inteiro, permitindo assim ao usuário visualizar, como redes de interação, grupos menores existentes ou realizar uma análise de enriquecimento do conjunto de genes de todo o genoma. Além dos sistemas de classificação já conhecidos como o *Gene Ontology* e o KEGG agora permite a utilização de outras duas opções: um sistema de mineração de dados que busca palavras chaves em resumos de artigos do Pub-Med e um sistema de clusterização hierarquica da própria rede do STRING que retorna clusters de tamanho entre 5 e 200 levando em consideração módulos que possuem muitas conexões. Na geração da rede existem configurações importantes que devem ser observadas para garantir a qualidade e confiança do resultado:

- ***Minimum required interaction score***: utiliza uma escala própria entre zero e um utilizando todas as associações verdadeiras nas evidências disponíveis.
- ***Textmining***: realiza uma análise estatística de co-citação em todos os resumos do Pub-Med assim como outras bases de dados.
- ***Experiments***: todas as interações existentes no banco de dados IMEx são novamente mapeadas e processadas retirando os dados duplicados. Os registros resultantes são aferidos contra o KEGG.
- ***Databases***: é baseado nas interações dos bancos de dados curados como KEGG, Reactome, BioCyc e Gene Ontology, assim como em bancos de dados legados como o PID e

BioCarta.

- **Co-expression:** a correlação entre genes é testada entre vários bancos de dados de expressão como por exemplo o NCBI *Gene Expression Omnibus*.
- **Neighborhood:** Através de comparação sistemáticas de genomas de procariotos verifica a semelhança entre os genes vizinhos.
- **Gene Fusion:** Assim como no *neighborhood*, utiliza a comparação com genomas de procariotos para identificação da fusão dos genes em cada organismo.
- **Co-ocurrence:** Também utiliza a comparação de genomas de procariotos para verificar a existência ou falta de proteínas que estão ligadas.

2.7.2 STITCH

STITCH é um banco de dados de interações proteína-pequenas moléculas que integra fontes experimentais e curadas manualmente com informações de mineração de texto (*textmining*) e previsões de interação (KUHNS et al., 2013). Atualmente possui cadastradas redes com 2031 organismos, 500 mil drogas, 9,6 milhões de proteínas e 1,6 bilhões de interações.

2.8 Redes Biológicas

Redes complexas de uma forma muito simplificada, corresponde ao conjunto de elementos que interagem entre si, são representadas por elementos matemáticos como os vértices (nós, nodos) e as arestas (interações, ligações, associações). No caso de redes sociais, cada pessoa representa um vértice e a amizade entre duas pessoas, é representada pela aresta. Uma rede complexa pode ser direcionada, quando existe uma orientação para a aresta, ou não-direcionada quando não há nenhuma orientação para a aresta.

No caso de redes biológicas (VERLI, 2014), a figura 2.13 (A) apresenta um exemplo de rede direcionada. O mecanismo de degradação ubiquitina-proteassoma: uma proteína destinada à degradação é marcada por ligações covalentes de moléculas de ubiquitina que transmitem um sinal para o proteassoma que, por sua vez, degrada a proteína em peptídeos. Essa representação é dita direcionada porque não é possível realizar o caminho inverso, partir dos peptídeos para formar a proteína via proteassoma. A figura 2.13 (B) apresenta um exemplo de rede não direcionada. A reação reversa de fosforilação e desfosforilação de adenosina difosfato. Te-

mos adenosina monofosfato (AMP) que recebe um grupo fosfato (PO_4) e se torna adenosina difosfato (ADP) que sua vez pode receber um grupo fosfato (PO_4) que se torna em adenosina trifosfato (ATP) que pode perder um grupo fosfato por vez até tornar-se AMP ($\text{AMP} \rightleftharpoons \text{ADP} \rightleftharpoons \text{ATP}$).

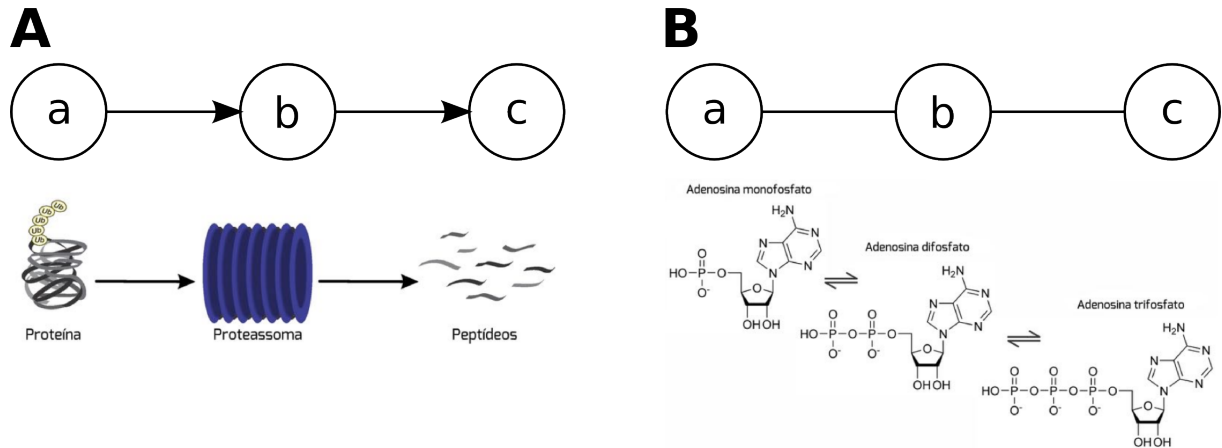


Figura 2.13: Exemplo de rede direcionada e não-direcionada. (A) Rede direcionada, a via de degradação de proteína; (B) Rede não-direcionada, a reação reversa de fosforilação de adenosina difosfato. Adaptado de (VERLI, 2014)

Pode-se definir uma rede complexa por meio de uma matriz de adjacência (A) como uma representação matricial de uma rede, onde cada elemento da matriz $a_{i,j}$ assume o valor 1(0) no caso de haver (ou não) uma aresta ligando os vértices identificados por i e j . Essa representação não traz informação sobre um possível direcionamento (i age sobre j mas j não age sobre i) nem sobre os pesos destas associações (figura 2.14).

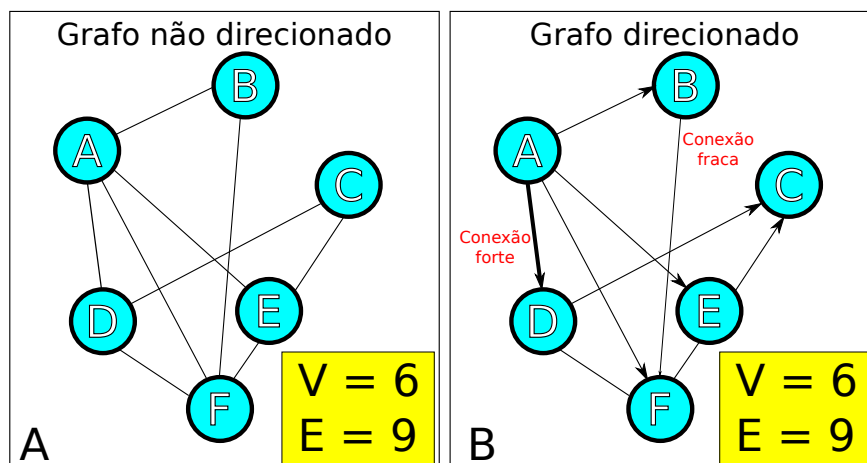


Figura 2.14: Exemplos de tipos de grafos. (A) Grafos não direcionados demonstram apenas as conexões entre os nós. (B) Grafos direcionados possuem informações sobre as conexões, direção e força (espessura da linha) em cada conexão. Fonte: O Autor, 2022.

2.8.1 Conectividade

A conectividade (k_i) de um vértice i de uma rede é definida como o número de arestas do vértice (i). A partir desta informação, é possível obter a conectividade para cada vértice via matriz de adjacência (equação 2.1).

$$k_i = \sum_j^N a_{ij} \quad (2.1)$$

Onde N é o número total de vértices, e a_{ij} é o elemento da matriz de adjacência (A). Através desta medida podemos definir também a conectividade média da rede $\langle k \rangle$ (equação 2.2):

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^N k_i. \quad (2.2)$$

2.8.2 Coeficiente de clusterização para redes não direcionadas

O coeficiente (C_i) de clusterização mensura a razão existente de arestas n entre os vizinhos de um dado vértice i e o número máximo possível destas arestas (equação 2.3):

$$C_i = \frac{2n}{k_i(k_i - 1)} = \frac{1}{k_i(k_i - 1)} \sum_{j=1}^N a_{ij} \sum_{m=1}^N a_{jm} a_{mi}. \quad (2.3)$$

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

Assim como na conectividade, podemos ter um valor do coeficiente de clusterização médio para uma rede (equação 2.4):

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^N C_i. \quad (2.4)$$

2.8.3 Modelos de Redes

2.8.3.1 Rede Aleatória

Em 1960, Erdős-Rényi (ERDŐS; RÉNYI, 1960) criaram um modelo de rede aleatória, iniciando a partir de uma rede com N vértices e considerando uma probabilidade p que um dado par de vértices seja conectado. Este processo gera uma rede com aproximadamente $\frac{pN(N-1)}{2}$ arestas distribuídas aleatoriamente como na figura 2.15 Aa. A distribuição da conectividade segue uma lei de distribuição binomial, que indica que muitos vértices terão a mesma quantidade

de arestas, enquanto que alta e baixa conectividades são raras como se pode observar pelo gráfico na figura 2.15 **Ab**. O coeficiente de clusterização é independente da conectividade do vértice, como é visto pelo gráfico $c(k)$ versus k , onde temos uma linha horizontal ($C(k) \approx p$), figura 2.15 **Ac**.

2.8.4 Rede Escala Livre

A rede escala livre é obtida por meio do modelo de crescimento de redes proposto por Barabási e colaboradores (BARABASI; OLTVAI, 2004) que, iniciando com um pequeno conjunto de vértices todos ligados uns aos outros, novos vértices são acrescentados um a um. As ligações de um novo vértice com os já existentes são criadas com uma probabilidade dada por:

$$p(k_i) = \frac{k_i}{\sum_{j=1}^N k_j} \langle k \rangle \quad (2.5)$$

onde k_i é a conectividade do vértice e $\langle k \rangle$ a conectividade média desejada. A rede é mostrada na figura 2.15 **Ba**. Esta rede apresenta uma distribuição de conectividade $p(k)$ na forma de uma lei de potência caracterizada por um expoente $\gamma = -3$, mostrada na figura 2.15 **Bb**. A probabilidade de existir vértices altamente conectados é estatisticamente mais significativa do que na rede aleatória e tais vértices são chamados de “hubs”.

A figura 2.15 **Bc**, mostra o comportamento do coeficiente de clusterização com respeito à conectividade. Vemos que é uma linha reta paralela ao eixo da conectividade. Isto significa que todos os vértices têm sempre um valor de coeficiente de clusterização muito próximo um dos outros, não importando a sua conectividade, logo não existe uma formação de módulo. No caso da 2.15 **Cc**, vemos que o coeficiente de clusterização tende a ser muito alto para vértices pouco conectados e baixo para vértices muito conectados (reta decrescente). Se tivéssemos uma curva gaussiana no lugar da reta decrescente na figura 2.15 **Bc**, poderíamos dizer que existe um tamanho preferencial de módulo, dado pela média da gaussiana.

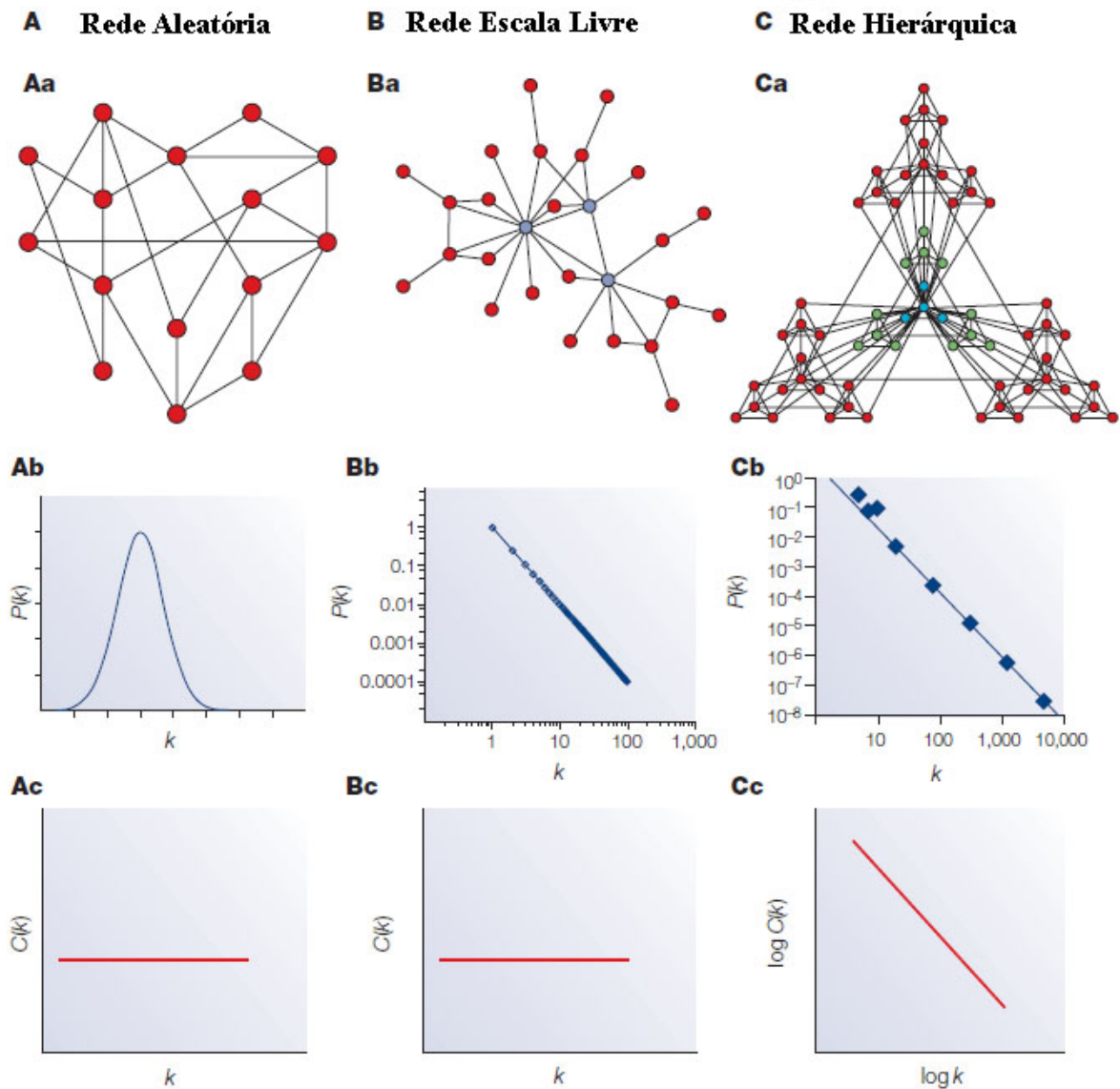


Figura 2.15: Principais modelos de redes, figura retirada do artigo do Barabási (BARABASI; OLTVAI, 2004). Esquemas para a **A** rede aleatória, **B** rede de escala livre e **C** rede hierárquica. Com as representações para suas redes topológicas (Aa, Ba e Ca), distribuição de conectividade (Ba, Bb e Bc) e coeficiente de clusterização (Ac, Bc e Cc).

2.8.5 Rede Hierárquica

Uma rede hierárquica (BARABASI; OLTVAI, 2004) pode ser construída partindo um bloco de N vértices todos ligados entre si, que é replicado m vezes. Uma rede aumentada é gerada ligando cada vértice central destes m módulos ao vértice central do módulo inicial, formando um super módulo de $(m+1)N$ vértices. Repetindo este processo *ad infinitum* obtemos uma rede hierárquica, como na figura 2.15 **Ca** que apresenta uma topologia livre de escala mas com uma estrutura modular. A distribuição de conectividades desta rede é uma de lei de potências com

expoente $\gamma = -2,26$, como na figura 2.15 **Cb**. O coeficiente de clusterização escala com uma lei de potência que segue $C(k) \approx k^\gamma$, como pode ser visto na figura 2.15 **Cc**.

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

2.8.6 Justificativa

A utilização desses métodos permitem uma visão sistêmica de toda a topologia da rede que, quando integrada às moléculas que fazem parte da estrutura e regulação, torna-se possível entender como o surgimento de doenças está relacionada a alterações na rede (CHARITOU; BRYAN; LYNN, 2016).

Rybarczyk e colaboradores desenvolveram uma metodologia e o transformaram em um pacote para a linguagem de programação R chamado transcriptograma, nessa metodologia os genes são ordenados em uma lista e através de informações geradas a partir de interações proteína-proteína com dados disponibilizados em bancos de dados públicos os correlaciona com a expressão gênica, todos os genes são analisados não sendo descartado nenhum, diferencialmente expressos ou não, pois genes que não são considerados diferencialmente expressos podem ter um efeito importante no sistema estudado ((RYBARCZYK-FILHO et al., 2011)).

Entre os principais repositórios de pacotes desenvolvidos para a linguagem de programação R está o Bioconductor (www.bioconductor.org). Ao submetermos um pacote para ser aceito no Bioconductor são realizadas diversas verificações e controles de qualidade no código e nos resultados gerados pelo pacote. Ao final das verificações são apresentados notas, avisos ou erros. Notas permitem que o pacote passe essa etapa e podem estar relacionados, por exemplo, a quantidade de linhas das funções do pacote. Avisos podem gerar problemas e pode estar relacionados, por exemplo, a quantidade de caracteres e indentação do código. Se o código possuir erros o pacote não é aceito pelo Biocondutor.

O transcriptograma em sua versão original, ao realizar as etapas de processamento, cria diretórios e arquivos de texto para organizar e permitir que o usuário acesse os resultados. O Bioconductor retorna erro durante a etapa de verificação do pacote caso sejam salvos dados em formato de arquivos nos discos de armazenamento.

Para resolvermos esse problema fizemos alterações no código do transcriptograma para que todos os resultados sejam salvos em um objeto do R. Salvar os resultados em um objeto do R facilita a manipulação dos dados pelo usuário, além de padronizar o formato e permitir a

construção de outros pacotes que expandam as funcionalidades do transcriptograma.

Além disso, todas as funções e processos foram inseridas em uma interface gráfica do usuário, em formato de *dashboard*, visando fornecer visualizações rápidas dos resultados obtidos e facilitar alterações nos parâmetros de execução.

3 *Objetivos*

Desenvolvimento de uma plataforma que permita a integração de dados (*Big Data*), tais como: redes biológicas, expressão gênica, processos biológicos, dentre outros. A plataforma deve ser de fácil acesso para diferentes níveis de usuários e que permita a integração de outras ferramentas oriundas de repositórios como o Bioconductor e R-cran.

3.1 **Objetivos específicos**

- Desenvolvimento de um pacote para projeção de expressão gênica sobre uma rede biológica.
- Criação de um objeto em R para salvar os dados gerados entre as etapas do processamento do transcriptograma.
- Integração da técnica de transcriptograma em ambiente R.
- Desenvolvimento de uma interface amigável para acessar as técnicas do transcriptograma.
- Validar o pacote utilizando dados de expressão gênica de amostras de pulmão de indivíduos fumantes, ex-fumantes e que nunca fumaram que tenham ou não adenocarcinoma em uma rede criada a partir de dados de interações proteína-proteína de *Homo sapiens*.

4 *Material e Métodos*

4.1 *Workflow*

Nesse trabalho, nós propomos a análise dos dados de expressão gênica de amostras de pulmão obtidas das tecnologias de sequenciamento microarranjo e Single-Cell RNA. Estes dados serão normalizados e avaliados aplicando o *workflow SimpleSingleCell* elaborado por (LUN; MCCARTHY; MARIONI, 2016). As amostras foram adquiridas dos estudos *PD-L1 amplification is associated with an immune cell rich phenotype in squamous cell cancer of the lung* (GOLDMANN et al., 2021) e *Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival* (LANDI et al., 2008). A tabela 4.1 possui os dados do título do artigo, autores, quantidade de amostras, tecnologia de sequenciamento utilizado e número de identificação no repositório *Gene Expression Omnibus*.

Tabela 4.1: Detalhes dos dados de expressão utilizados na análise

Artigo	Autores	Quantidade de amostras	Tecnologia de sequenciamento	GSE
PD-L1 amplification is associated with an immune cell rich phenotype in squamous cell cancer of the lung.	GOLDMANN et al.	218	RNA-seq, Illumina HiSeq 2500	GSE81089
Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival	LANDI et al.	135	Microarray (Affymetrix)	GSE10072

Fonte: O autor (2022).

Paralelamente iremos adquirir dados para criação da rede biológica utilizando o STRING. Serão utilizados os seguintes parâmetros: Experimentos e *database*, com *score* de 0.80 e será realizado o *download* de todas as interações de *Homo sapiens* (taxon id = 9606). A integração desses dados serão analisados usando a metodologia do transcriptograma (figura 4.1).

4.2 *levi*

O *levi* será desenvolvido utilizando algumas linguagens de programação de forma híbrida. Abaixo, temos os conceitos das ferramentas utilizadas e as explicações dos cálculos para con-

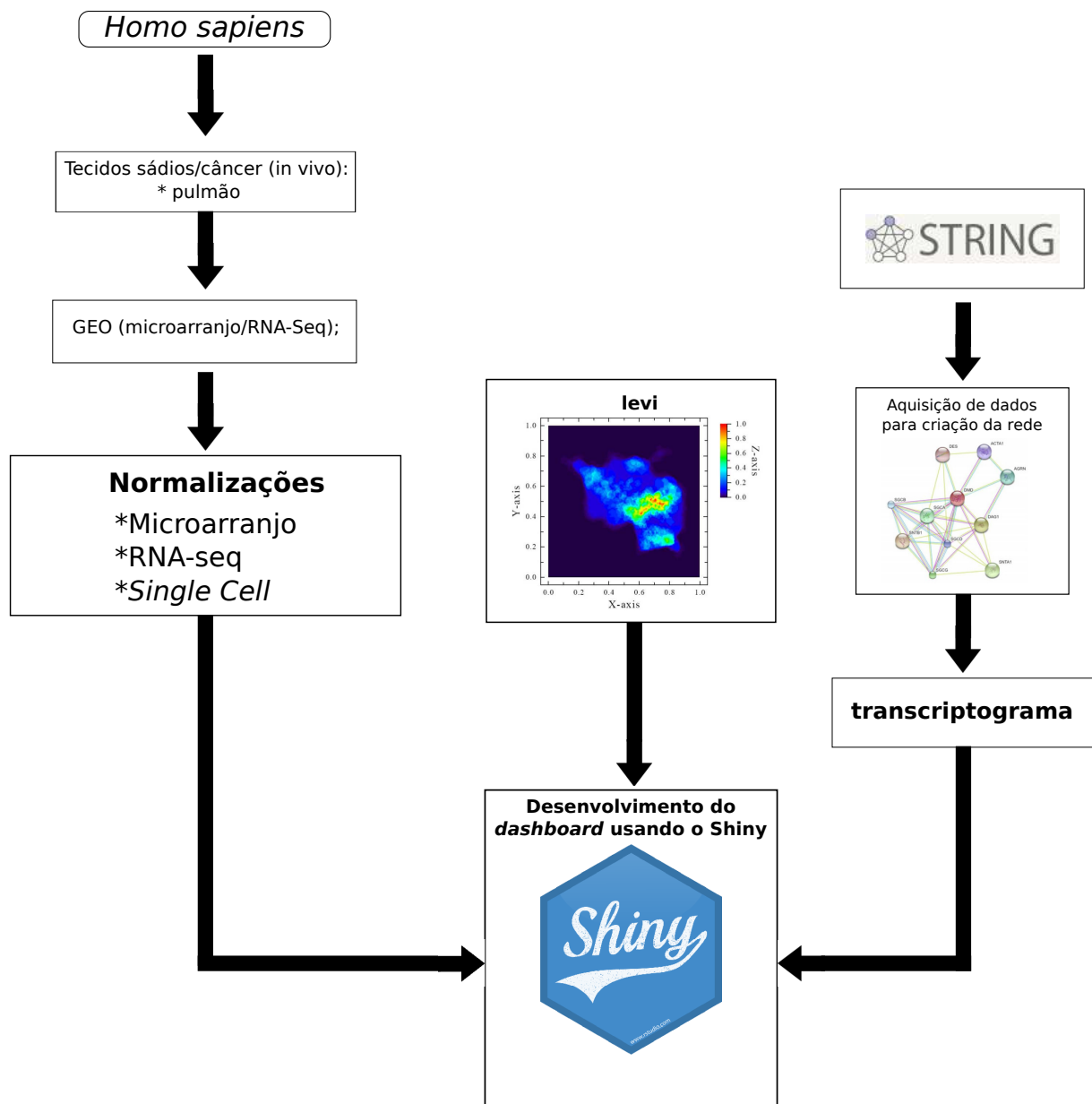


Figura 4.1: *Workflow* com as etapas do projeto. O projeto será dividido em três etapas, sendo as duas primeiras executadas paralelamente. (A) Teremos a obtenção dos dados para análise de expressão gênica de *Homo sapiens*. (B) Ao mesmo tempo faremos a aquisição de dados em bancos de dados *online* para criação das redes de interação. (C) Com a utilização dos métodos do transcriptograma e o *levi* iremos desenvolver uma ferramenta em *dashboard* para facilitar a utilização das ferramentas de análise de rede.

traste, resolução, zoom e suavização que são responsáveis por alterações na geração de imagem do *landscape*.

4.2.1 R

R é uma linguagem e um ambiente de desenvolvimento voltado principalmente para computação estatística (inferência, simulações, *data mining*, etc) e gráficos, está disponível como um *software* livre, atualmente está na versão 4.2.2 (publicada em 31/10/2022) e pode ser executado em sistemas operacionais Windows, Linux, MacOS (R Core Team, 2018).

Entre as facilidades da utilização da linguagem R, podemos citar:

- Facilita o armazenamento e manipulação de dados;
- Possui variadas funções e pacotes para manipulação de vetores e matrizes;
- Pacotes e ferramentas que permitem a análise dos dados;
- Integração com outras linguagens de programação;
- Possibilidade de criação de pacotes com interface gráfica.

O ambiente de desenvolvimento integrado (IDE) utilizado foi o RStudio. O RStudio possui uma interface gráfica do usuário que permite a utilização de botões, navegação entre os diretórios, leitura da documentação dos pacotes e facilita a codificação de scripts em R.

O R foi usado em 85% do desenvolvimento e é responsável por receber os parâmetros do usuário como: arquivo com dados dos nodos, arquivo com valores de expressão, nomes das colunas dos dados de expressão que serão utilizadas para o processamento, coluna com o Gene Symbol; criação de funções utilizadas internamente no processamento; execução das linhas de código; desenvolvimento da interface de usuário com o Shiny e geração dos resultados.

Um paradigma de programação bastante utilizado é a programação orientada a objetos (POO) ela consiste na utilização do conceito de objetos para armazenar atributos (dados) e métodos (códigos).

A linguagem R propicia a utilização de programação orientada a objetos para a criação de programas. Para o desenvolvimento do presente trabalho algumas classes de objetos foram importantes como os vetores, matrizes e listas.

- **Vetores** - variável unidimensional de tamanho fixo. Ao ser criado um vetor é informado quantos elementos poderá ter, esse valor não poderá ser alterado posteriormente.
- **Listas** - variável unidimensional de tamanho dinâmico. É possível aumentar ou diminuir o seu tamanho conforme novos elementos são incluídos ou excluídos.

4.2.2 C++

O C++ é uma das linguagens de programação mais utilizadas no mundo, seu uso vai desde a criação de sistemas operacionais até o desenvolvimento de programas para serem utilizados em redes de computadores e aplicações cliente/servidor. Foi criada no início da década de 1980 por Bjarne Stroustrup e uma das suas principais características é a programação orientada a objetos (DEITEL; DEITEL, 1999).

A linguagem R possui limitação na velocidade de execução de laços de repetição, para resolver esse problema foi utilizada a linguagem de programação C++, que faz parte de 14.5% de todo o código, e é responsável por executar alguns laços de repetição com maior velocidade.

4.2.3 CSS

O *Cascading Style Sheets* (CSS) permite que o programador defina regras de estilos que são aplicadas aos elementos das páginas HTML. Essas regras podem ser aplicadas a um elemento único, a um grupo de elementos ou a um tipo específico de elemento. As regras de estilos permitem alterar a renderização dos elementos como cores, alinhamentos e bordas facilitando a execução na maioria dos navegadores (GOODMAN, 2002).

O CSS foi utilizado em 0.5% do código e trabalha em conjunto com o Shiny para seleção de cores e efeitos na barra de carregamento que aparece quando o programa está em execução no seu modo gráfico.

4.2.4 Equações

Contraste

Aplica o cálculo para contraste no landscape. O valor de contraste altera a quantia de bits por pixel.

$$\text{contraste} = 0.1 - \left(0.1 \left(\frac{x}{100} \right) \right) \quad (4.1)$$

Para um valor de entrada, na formula sendo representado pela variável x, de “1” no contraste é aplicado o resultado do cálculo de 0.099 na geração da imagem, enquanto um valor de entrada de contraste de “100” aplica o resultado de 0.001 na geração da imagem.

O valor de contraste está relacionado ao tamanho da visualização de cada vertice na geração

do *landscape*. Um valor alto para esse parâmetro irá aumentar o raio para cada vértice, enquanto que um valor menor irá diminuir o raio para cada vértice.

Resolução

Altera a dimensão do *landscape*. Está relacionado ao tamanho da matriz para criação da imagem. Cada célula da matriz é transformada em um pixel para a geração do *landscape*.

$$\text{resolução} = \text{inteiro}\left(\left(\frac{y}{100}\right)210 + 30\right) \quad (4.2)$$

A equação retorna o valor inteiro do cálculo. Um valor de entrada, na formula sendo representado pela variável y , igual a “1” gera uma matriz de 32x32, da mesma forma um valor de entrada igual a “100” gera uma matriz de 240x240. Esses valores são respectivamente o menor e maior valor possível para a resolução no *levi*.

Uma resolução maior fará com que o *landscape* possua uma maior quantidade de pixels, com isso melhora a visualização do resultado gerado, contudo, quanto maior a resolução, proporcionalmente, também é maior o tempo de processamento.

Zoom

Aumenta ou diminui a escala de visualização da rede do *landscape*. O algoritmo utilizado para gerar a imagem sempre cria o *landscape* de forma centralizada a partir da célula existente no ponto médio dos eixos x e y da matriz. A variável z representa a entrada do usuário e pode receber valores entre 0 e 100.

$$\text{zoom} = 0.2\left(\frac{z}{100}\right) - 0.2 \quad (4.3)$$

Para o zoom temos como exemplo que o valor de entrada “0” aplica o resultado de -0.2 no *landscape*. O valor de entrada igual a “100” aplica o resultado de 0 no *landscape*.

Suavização

Aplica o cálculo de suavização na imagem gerada do *landscape*.

$$\text{suavização} = \text{inteiro}\left(\frac{w}{100}\right)18 \quad (4.4)$$

Se a suavização, na formula sendo representado pela variável w , for menor ou igual a zero ela recebe 1. Com isso o menor e maior valor resultantes do cálculo são respectivamente 1 e 18.

O parâmetro de suavização altera os valores do vértice virtual no ponto médio entre a conexão de dois vértices reais. O que temos na prática é que se o valor de suavização for baixo o landscape gerado possuirá, na sua visualização, pixels com intensidades de expressão diferentes, se a rede for grande, esse tipo de visualização pode dificultar a leitura de áreas de expressão.

A figura 4.2 exemplifica o workflow para geração do landscape pelo levi. (A) temos uma rede hipotética criada utilizando o Medusa na sua versão 3.0 essa rede é utilizada como entrada para o levi. (B) ao executar o levi são verificados os valores para os parâmetros de contraste, resolução, zoom e suavização e realizado os cálculos para cada uma dessas configurações. (C) a primeira etapa para geração do landscape é a criação do contorno utilizando os dados dos nodos e interações, em conjunto com os parâmetros definidos pelo usuário. O tamanho da matriz é definido pelo valor de resolução. Os valores de constraste e zoom definem se as células da matriz recebem o valor 0 ou 10. Células com valor 10 são usadas para definir o contorno do landscape. (D) em seguida são aplicados os valores de expressão e suavização substituindo os valores das células que estavam com valor 10. (E) por fim, o landscape é gerado.

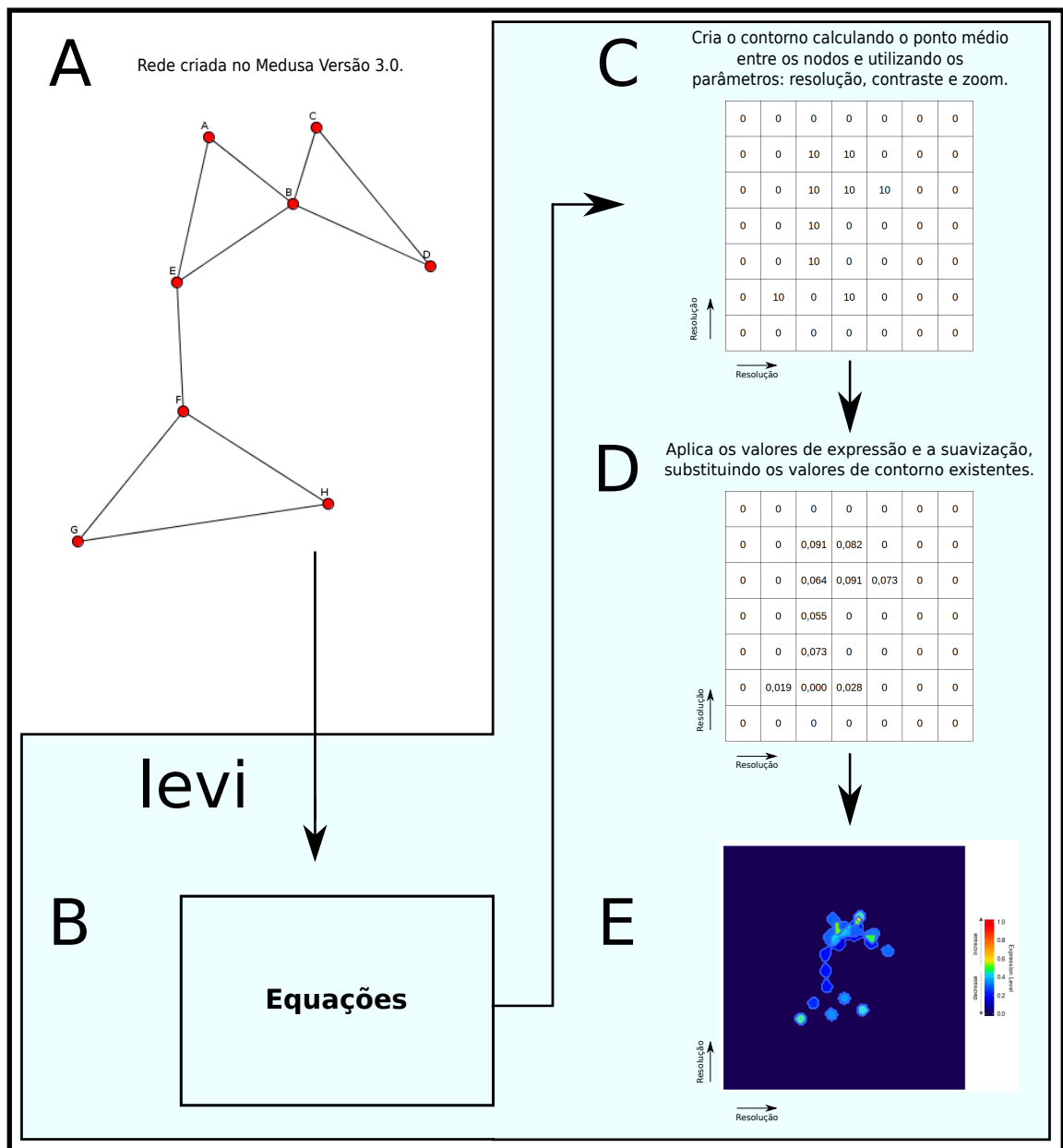


Figura 4.2: Workflow demonstrando as etapas de geração do landscape pelo levi utilizando o padrão (= 50) para os valores de resolução, contraste, suavização e zoom.

4.3 *transcriptograma*

O transcriptograma é um programa desenvolvido para o ambiente de programação R que utiliza como entrada uma rede e realiza análises utilizando o método de Monte Carlo para identificação de clusters ¹. Ele executa três funções principais: ordenamento, modularidade e análise da expressão gênica.

O transcriptograma salva no computador onde foi executado os resultados dos seus processos em arquivos no formato de texto. Nesse trabalho modificamos o transcriptograma para que todos os resultados entre as etapas de processamento sejam salvas em um objeto do R visando atender as exigências para ser aceito no Bioconductor e principalmente filtrar seu resultado em variáveis para facilitar a leitura do *dashboard* criado em Shiny para visualização dos resultados de forma interativa. Salvar os resultados das etapas em uma objeto do R facilita a manipulação para geração de novas análises, visualizações ou integração com outros pacotes do R.

Com os genes ordenados pelo transcriptograma é possível utilizar a ferramenta *levi* para a geração panorâmica da rede facilitando a visualização das regiões de interesse.

4.4 Método de ordenamento de redes em uma dimensão

Partindo de uma rede de interações proteína-proteína, não direcionadas, sem pesos nas interações, podemos transformá-la numa matriz de adjacência da seguinte maneira. Seja G uma rede com N vértices (proteínas) e E arestas (interações), podemos rotular de maneira arbitrária os vértices com números inteiros no intervalo $[1, N]$, onde a matriz de adjacência A tem N colunas e N linhas, de tal maneira que cada elemento a_{ij} pode ter valor lógico FALSO ou VERDADEIRO. Atribuímos valor VERDADEIRO para o caso dos vértices i e j interagirem entre si, caso contrário o valor será FALSO (figura 4.3).

¹ *cluster* é um neologismo, ele tem sentido de aglomeração, agrupamento.

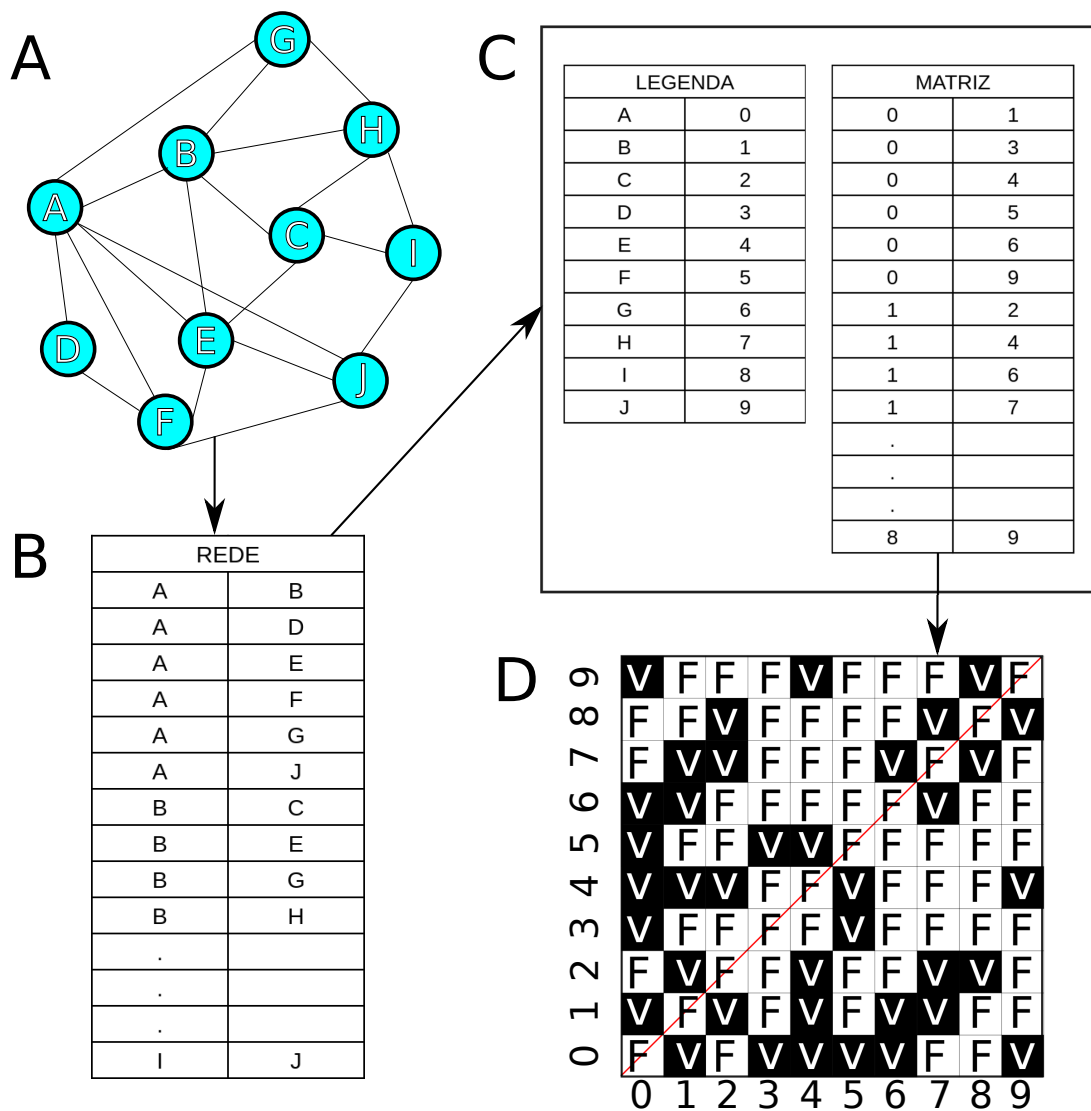


Figura 4.3: Transformação da rede em matriz de adjacência. **(A)** Rede hipotética de 10 vértices e 20 arestas. **(B)** A rede convertida em forma de lista. **(C)** Transformação da rede em formato de lista para o formato numérico para ser convertida em matriz. **(D)** Rede no formato de matriz de adjacência, onde V significa que existe interação entre os elementos e F significa que não existe interação entre os elementos.

A ordem de numeração dos vértices segue de forma crescente para cada novo vértice identificado, porém, essa ordem pode ser alterada por meio de permutações de linhas e colunas. Para cada alteração, uma nova matriz de adjacência é criada, sendo assim, o número possível de matrizes de adjacência criadas é $N!$ (fatorial de N). Por exemplo, uma rede com 20 vértices terá $20!$ (2.43×10^{18}) matrizes de adjacência para a mesma rede. Neste exemplo, sabemos que todas as matrizes representam a rede, algumas dessas matrizes podem evidenciar possíveis agrupamentos de vértices que podem ser relevantes.

Uma maneira de encontrar as matrizes que possam ser relevantes é arrancar os vértices aproximando os elementos VERDADEIROS da diagonal da matriz e afastar os elementos FALSOS,

como apresentado na figura 4.4.

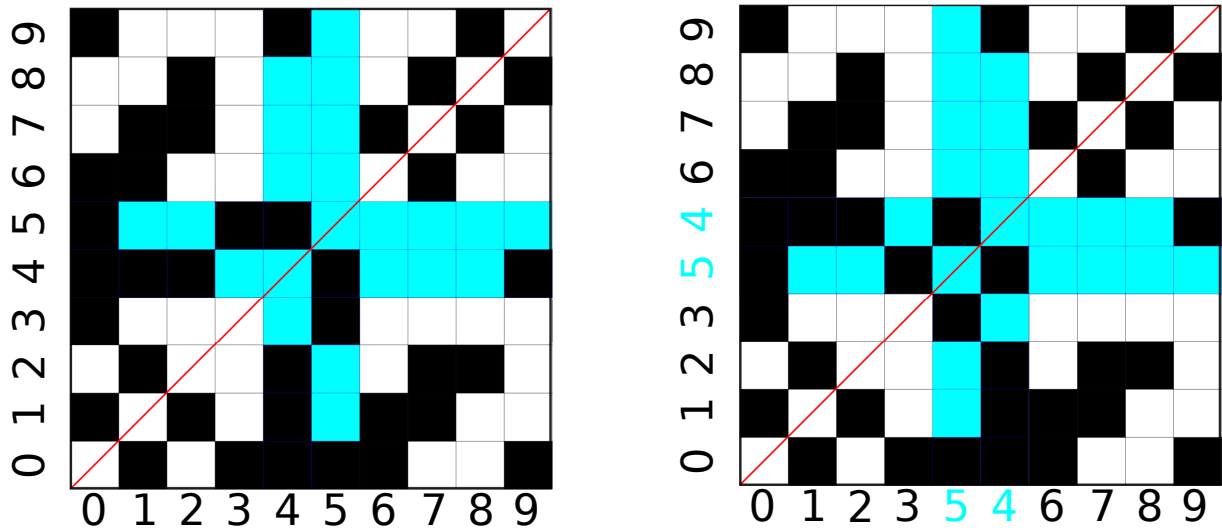


Figura 4.4: (A) Matriz de adjacência com 10 vértices numerados de 0 até 9. (B) Trata-se da mesma matriz apresentada em (A), porém foram realizadas duas trocas de linhas e colunas da mesma (coluna 4 foi trocada pela coluna 5, o mesmo foi realizado para as linhas).

A matriz B apresenta mais elementos próximos da diagonal do que a matriz A, logo a matriz B é o resultado desejado. Para alcançar este objetivo são necessárias duas condições:

- **Condição 1:** A melhor distribuição a ser atingida é aquela onde teremos menos interfaces VERDADEIRO/FALSO, ou seja, o melhor resultado é ter um grande número de elementos VERDADEIRO próximos (figura 4.5).
- **Condição 2:** A distância dos elementos VERDADEIROS em relação à diagonal da matriz de adjacência é um importante fator de aproximação dos elementos. Quanto mais elementos próximos da diagonal, menos elementos estarão presentes nas extremidades da matriz, assim facilitando a visualização de agrupamentos (figura 4.6).

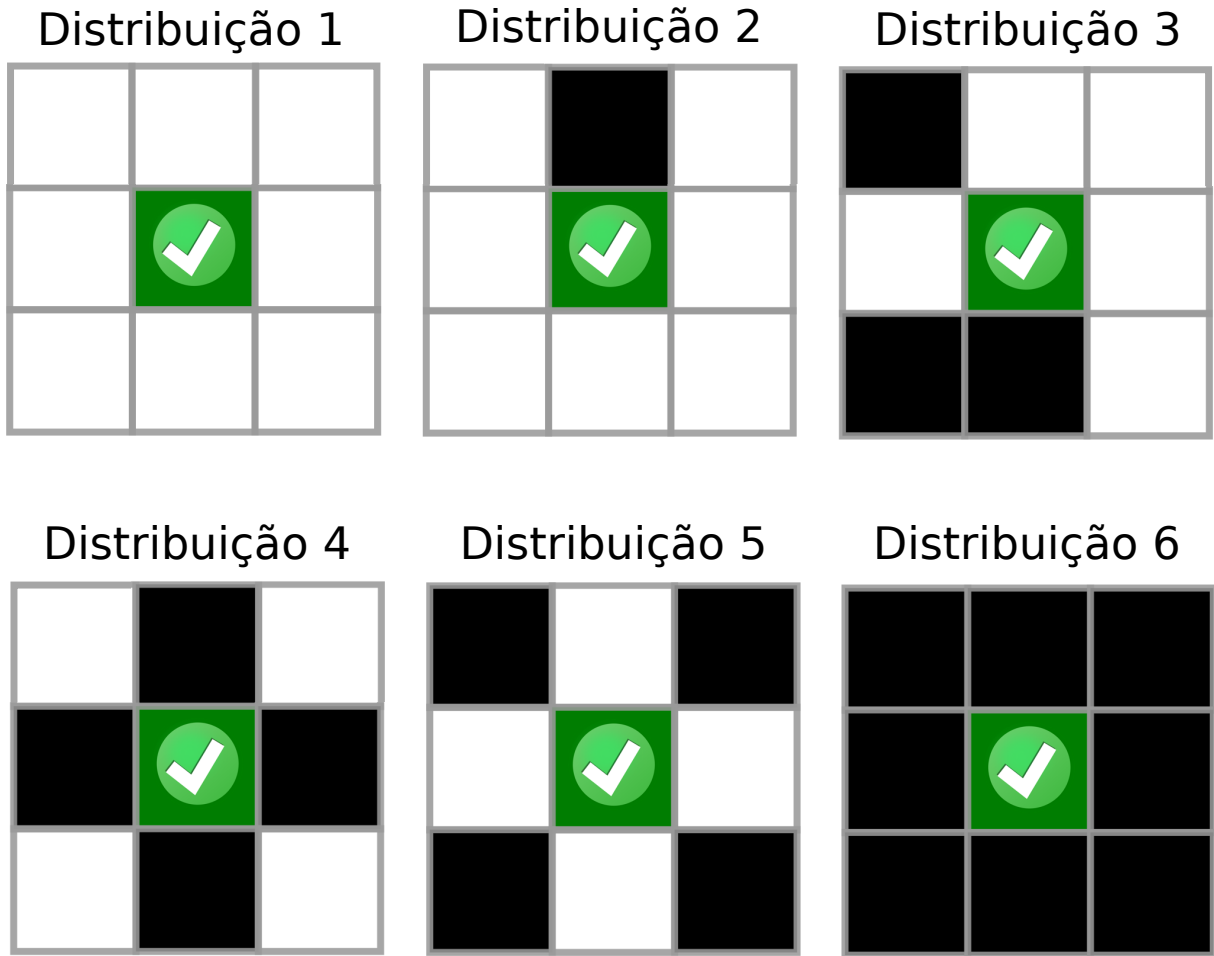


Figura 4.5: Proximidade dos vizinhos em relação ao elemento central na matriz de adjacência. A figura mostra seis possibilidades de vizinhos do elemento central. A melhor distribuição é aquela que apresenta maior número de VERDADEIROS em torno do elemento central, no caso a distribuição nº 6 é a melhor distribuição dentre todas as possibilidades.

4.4.1 Modelo Cruz

No artigo de Rybarczyk-Filho e colaboradores (RYBARCZYK-FILHO et al., 2011) foi apresentado um modelo que atende a condição 1 e 2 da seção 4.4. O modelo avalia os vizinhos acima, abaixo, à esquerda e à direita de um elemento VERDADEIRO na matriz de adjacência (figura 4.7) multiplica pela distância euclidiana do elemento em relação a diagonal da matriz de adjacência (Equação 4.5).

$$\varepsilon = \sum_{j=1}^V \sum_{i=1}^V d_{i,j} \{ |a_{i,j} - a_{i+1,j}| + |a_{i,j} - a_{i-1,j}| + |a_{i,j} - a_{i,j+1}| + |a_{i,j} - a_{i,j-1}| \} \quad (4.5)$$

Este cálculo é realizado para todos os elementos VERDADEIROS da matriz e somam-se seus resultados. O resultado final é chamado de “custo energético” (ε) da matriz de adjacência. Aplica-se o método de Monte Carlo para minimizar o ε da matriz.

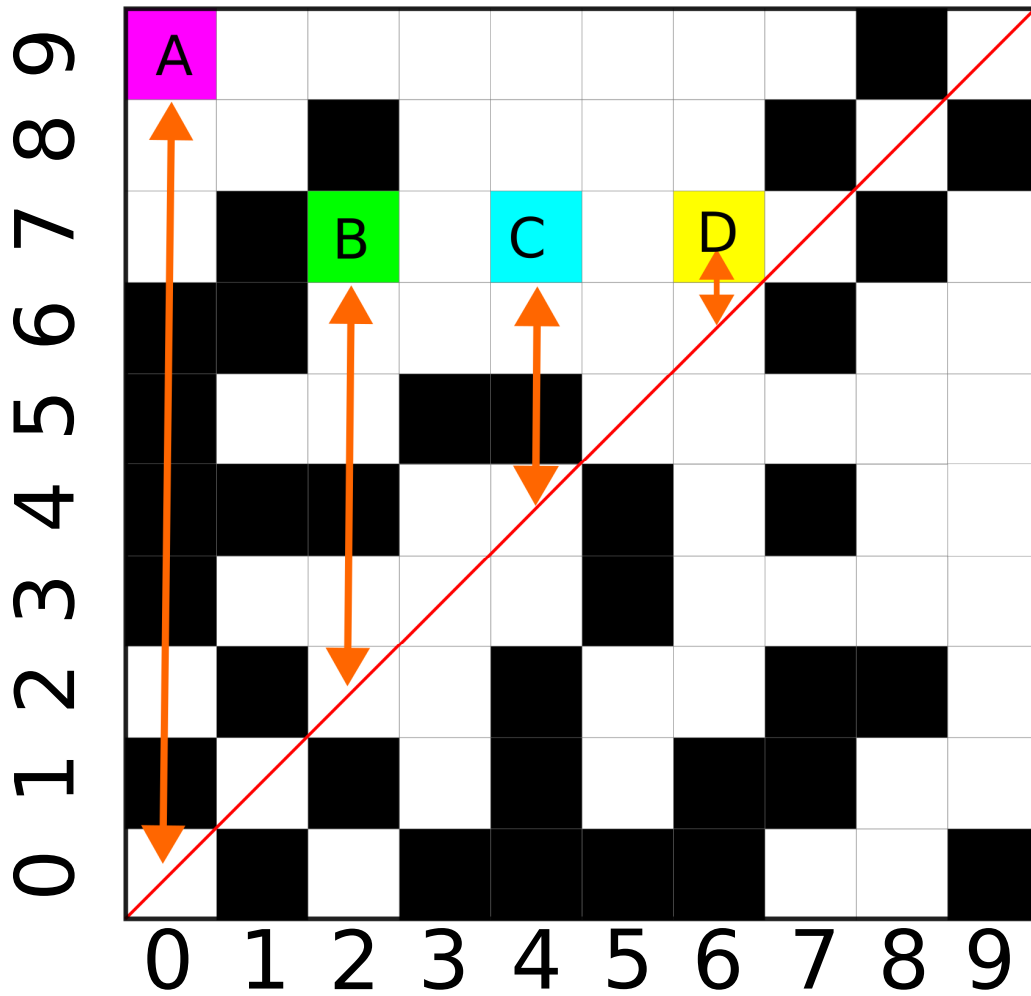


Figura 4.6: Representação de quatro possíveis distâncias do elemento central em relação a diagonal principal de uma matriz de adjacência. A diagonal principal é representada pela linha vermelha na matriz. A distância até o elemento A é a distância menos favorável, pois o elemento central está muito afastado em relação a diagonal. A até o elemento D pode ser considerada a distância mais favorável em relação elemento central e diagonal. O caso ideal é que a distância entre o elemento central e a diagonal seja o menor possível.

A metodologia se baseia na aleatorização dos vértices da matriz A. Realizam-se permutações de linhas e as respectivas colunas, ou seja, nenhuma interação é perdida ou criada durante o processo, a cada permutação é calculada a função “custo energético” (ϵ). Comparam-se os valores de ϵ_i antes e depois da permutação ϵ_f . Verifica-se $\epsilon_f < \epsilon_i$, caso seja, a nova distribuição da matriz de adjacência é aceita e realiza-se novas permutações. Caso $\epsilon_f > \epsilon_i$, a distribuição poderá ser aceita com uma probabilidade de $\exp - \frac{(\epsilon_f - \epsilon_i)}{T}$, onde T é um parâmetro análogo a temperatura no Método de Monte Carlo. Caso ϵ_f não seja aceito, então a matriz de adjacência retorna para o estado de ϵ_i .

Para a simulação, T inicia com um valor T_0 que é um valor igual ao “custo energético” inicial da matriz de adjacência ϵ_0 . Esse valor de T é diminuído gradativamente ao longo da

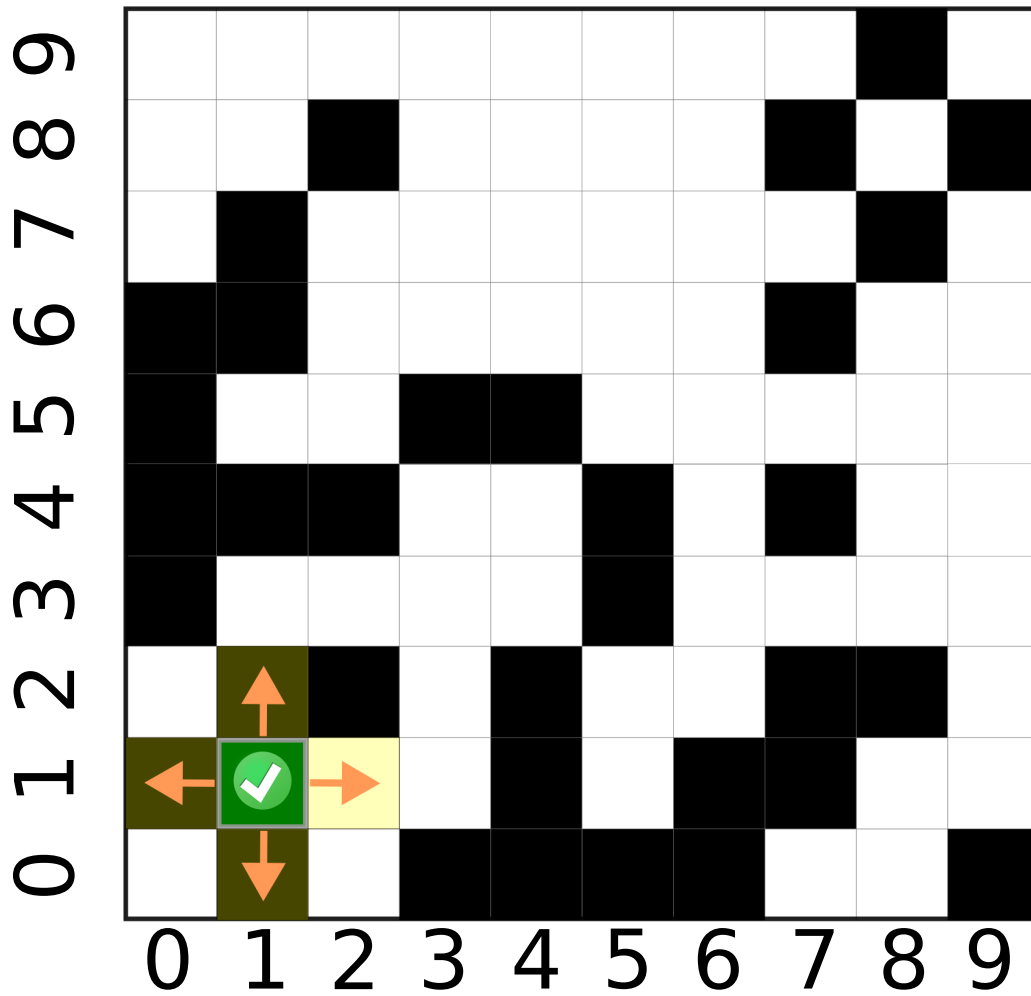


Figura 4.7: Avaliação da vizinhança do elemento central no modelo “cruz”, onde o elemento central é representado pelo símbolo de confirmação verde e branco e os vizinhos avaliados são representados por quadrados amarelos com setas nas direções que são avaliados os elementos em seu entorno (conforme equação 4.5).

simulação, espera-se um tempo suficiente para que o sistema atinja o equilíbrio termodinâmico para a temperatura em questão. A temperatura é reajustada por meio de uma razão de arrefecimento μ , tal que $T^* = \mu T$, sendo $0 < \mu < 1$, onde T^* é a nova temperatura. Este processo é repetido até a temperatura seja próxima de zero. Esta técnica é conhecida como *simulated annealing*, muito empregada em simulações para minimização de energia (figura 4.8).

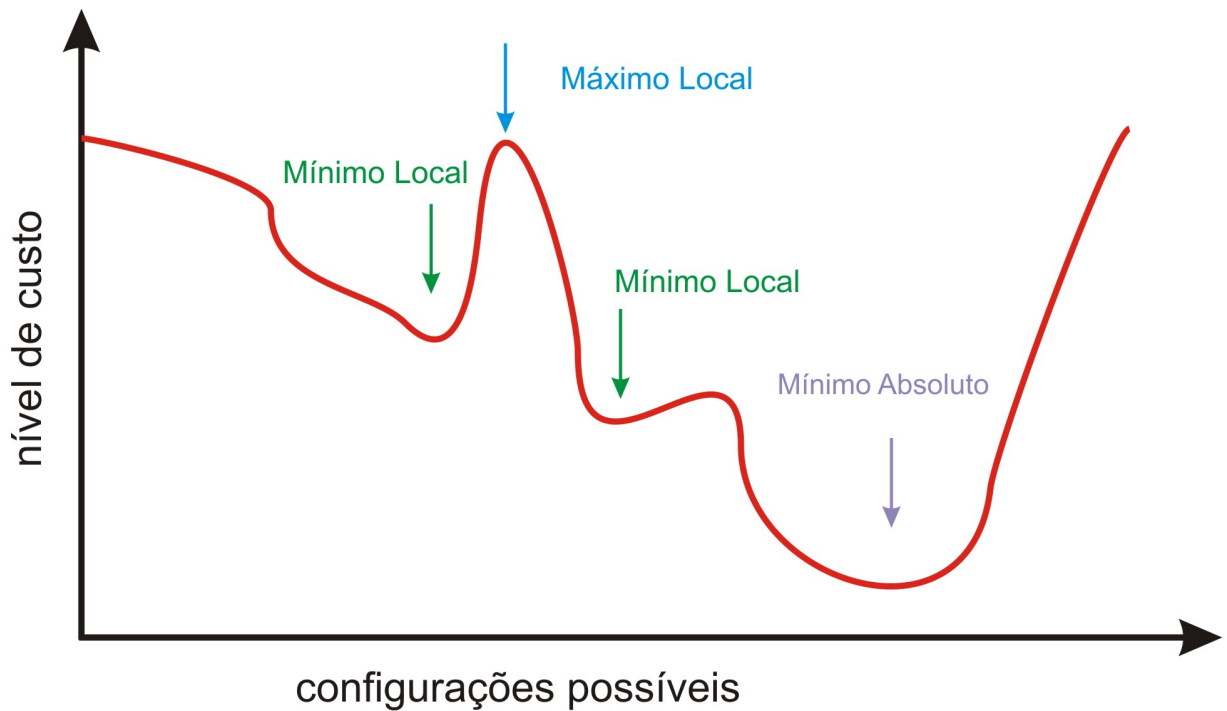


Figura 4.8: Gráfico de nível de custo energético por configurações possíveis de uma matriz de adjacência. Este gráfico representa a aplicação do *simulated annealing* em uma matriz hipotética para todas as suas possíveis configurações. O objetivo da técnica é atingir o **Mínimo Absoluto** passando pelos **Máximos Locais** sem que a simulação termine em algum **Mínimo Local**.

4.4.2 Modelo X

Este modelo foi proposto por Molan-Rybarczyk-Filho (MOLAN; RYBARCZYK-FILHO, 2014) a partir de uma alteração na equação do modelo cruz para o cálculo de proximidade dos elementos vizinhos em relação ao elemento central (equação 4.6). O Modelo X consiste numa avaliação de vizinhos às diagonais do elemento central: superior direito, superior esquerdo, inferior direito e inferior esquerdo (figura 4.9).

$$\varepsilon = \sum_{j=1}^V \sum_{i=1}^V d_{i,j} \{ |a_{i,j} - a_{i-1,j+1}| + |a_{i,j} - a_{i-1,j-1}| + |a_{i,j} - a_{i+1,j+1}| + |a_{i,j} - a_{i+1,j-1}| \} \quad (4.6)$$

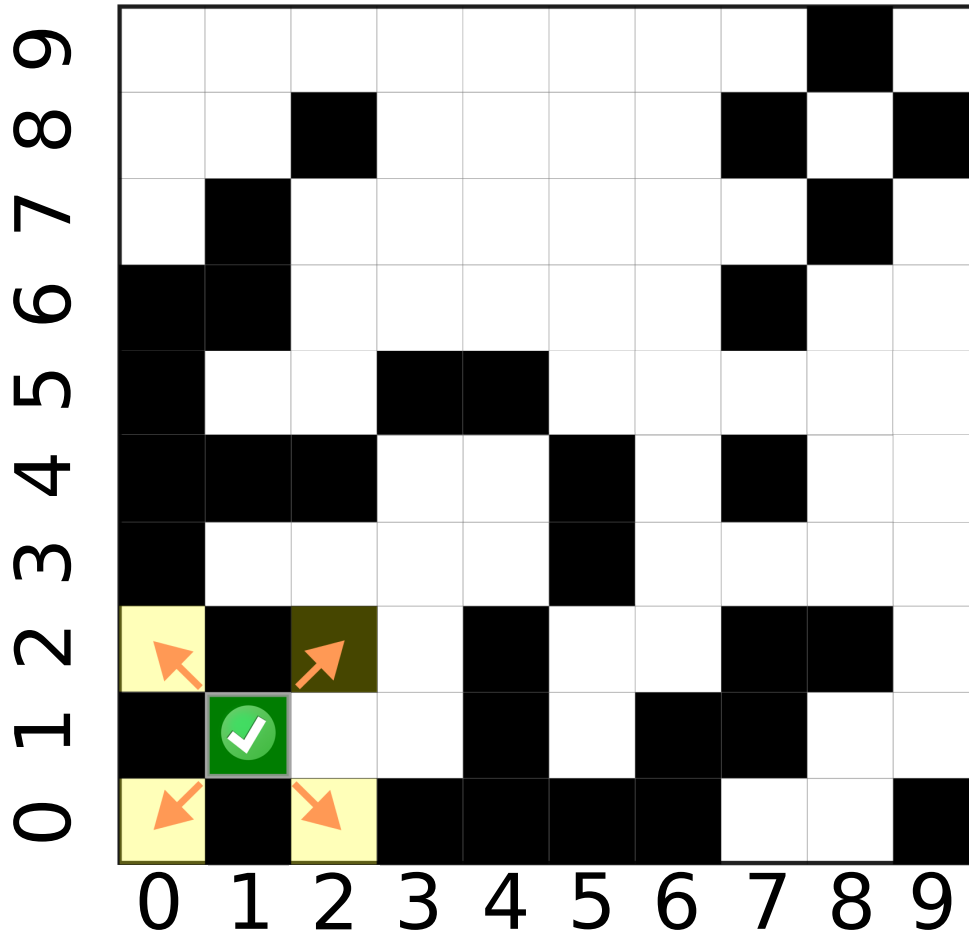


Figura 4.9: Avaliação da vizinhança do elemento central no modelo “X”, onde o elemento central é representado pelo símbolo de confirmação verde e branco e os vizinhos avaliados são representados por quadrados amarelos com setas nas direções que são avaliados os elementos em seu entorno (conforme equação 4.6).

4.4.3 Modelo Anel

Este modelo foi proposto por Molan-Rybarczyk-Filho (MOLAN; RYBARCZYK-FILHO, 2014) a partir da união dos modelos cruz e X para o cálculo de proximidade dos elementos vizinhos em relação ao elemento central (equação 4.7). O Modelo Anel consiste numa avaliação de todos os vizinhos do elemento central (figura 4.10).

$$\varepsilon = \sum_{j=1}^V \sum_{i=1}^V d_{i,j} \{ |a_{i,j} - a_{i-1,j-1}| + |a_{i,j} - a_{i-1,j}| + |a_{i,j} - a_{i-1,j+1}| + |a_{i,j} - a_{i,j+1}| + |a_{i,j} - a_{i+1,j+1}| + |a_{i,j} - a_{i+1,j}| + |a_{i,j} - a_{i+1,j-1}| + |a_{i,j} - a_{i,j-1}| \} \quad (4.7)$$

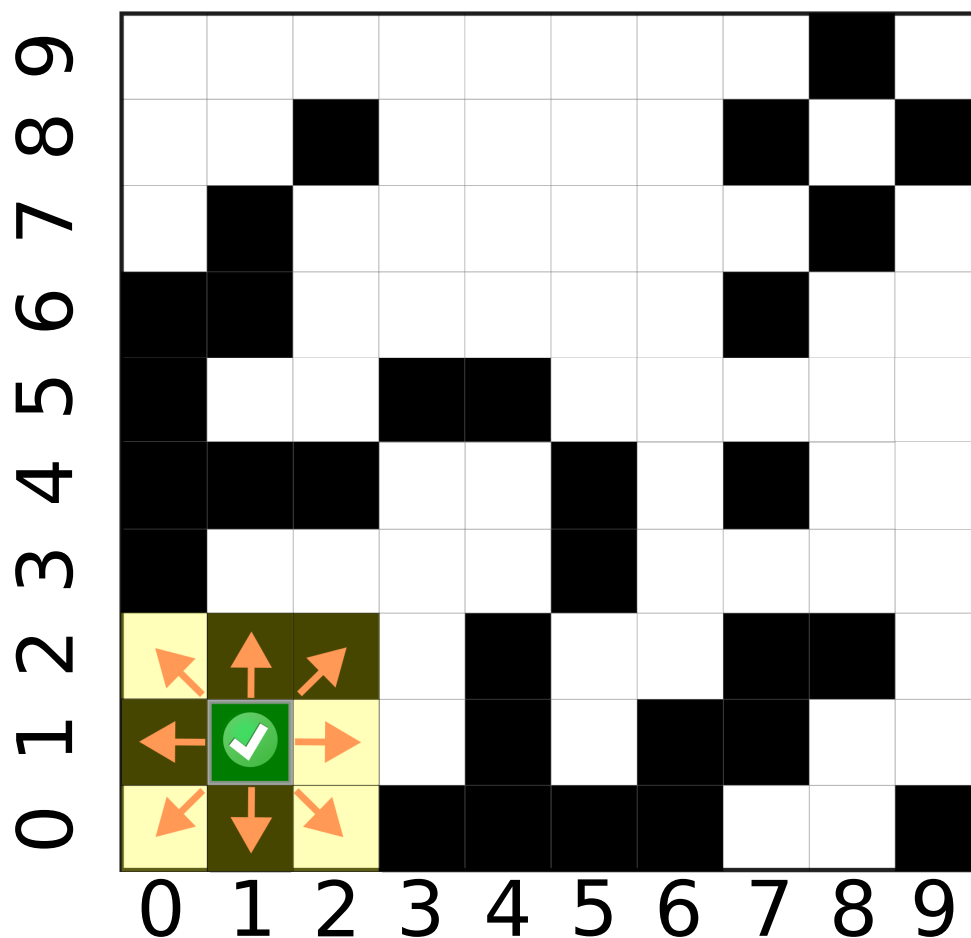


Figura 4.10: Avaliação da vizinhança do elemento central no modelo “Anel”, onde o elemento central é representado pelo símbolo de confirmação verde e branco e os vizinhos avaliados são representados por quadrados amarelos com setas nas direções que são avaliados os elementos em seu entorno (conforme equação 4.7).

4.5 Modularidade

A definição de um módulo funcional é um conjunto de elementos (proteínas, genes, etc) que interagem entre si de alguma forma (FRASER, 2005; VINOGRADOV, 2008) e realizam alguma função específica, como, por exemplo, o ciclo do ácido tricarboxílico ou ciclo de Krebs, que possui uma série reações catabólicas e anabólicas com a função de oxidar a acetil-CoA (NELSON; COX, 2004). Para o cálculo da modularidade, não consideramos as classificações de natureza biológica (ontologias, rotas metabólicas, etc) para evidenciar os módulos. Basicamente, os módulos são o resultado do rearranjo da matriz de adjacência.

Para a visualização dos módulos de uma maneira mais simples, utiliza-se uma função que mensura o número de interações entre os vértices contidos em uma janela de largura w dividido pela quantidade de conexões que cada vértice possui dentro e fora da janela w e atribui-se o valor para o vértice central da janela. A janela percorre todos os vértices da rede com condições de

contorno periódica. A figura 4.11 apresenta um exemplo do cálculo da modularidade para uma rede com dez vértices previamente ordenado pelo método (CFM), onde aplica-se uma janela de largura igual a três ($w = 3$).

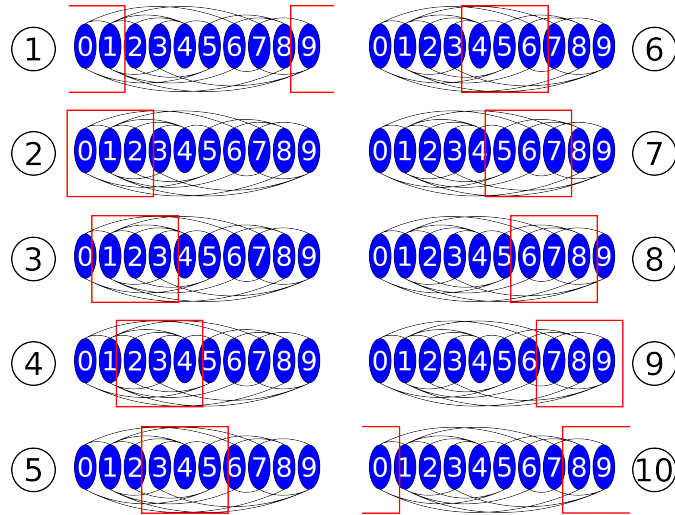


Figura 4.11: A figura apresenta 10 janelas para calcular a modularidade de uma rede com 10 vértices. Para calcular a modularidade utiliza-se uma função que mensura o número de interações entre os vértices contidos em uma janela de largura w dividido pela quantidade de conexões que cada vértice possui dentro e fora da janela w e atribui-se o valor para o vértice central da janela. A janela percorre toda a rede com condições de contorno periódicas. Fonte: (BIAZOTTI, 2016)

A figura 4.12 mostra a formação dos agrupamentos dos vértices baseada no cálculo de modularidade. Vemos dois grandes agrupamentos, um aproximadamente 0.33 de modularidade e outro com no máximo 0.40 de modularidade. Entre estes grupos o nível máximo de modularidade é aproximadamente 0.1. Podemos interpretar da seguinte maneira: No agrupamento com modularidade 0.4, os vértices dentro da janela compartilham no máximo 40% de suas interações dentro da janela e os mesmos vértices têm 60% de suas interações fora com vértices fora da janela. O ideal seria atingir o valor 1.0, que significa que todos os vértices dentro da janela compartilham entre si todas as suas interações.

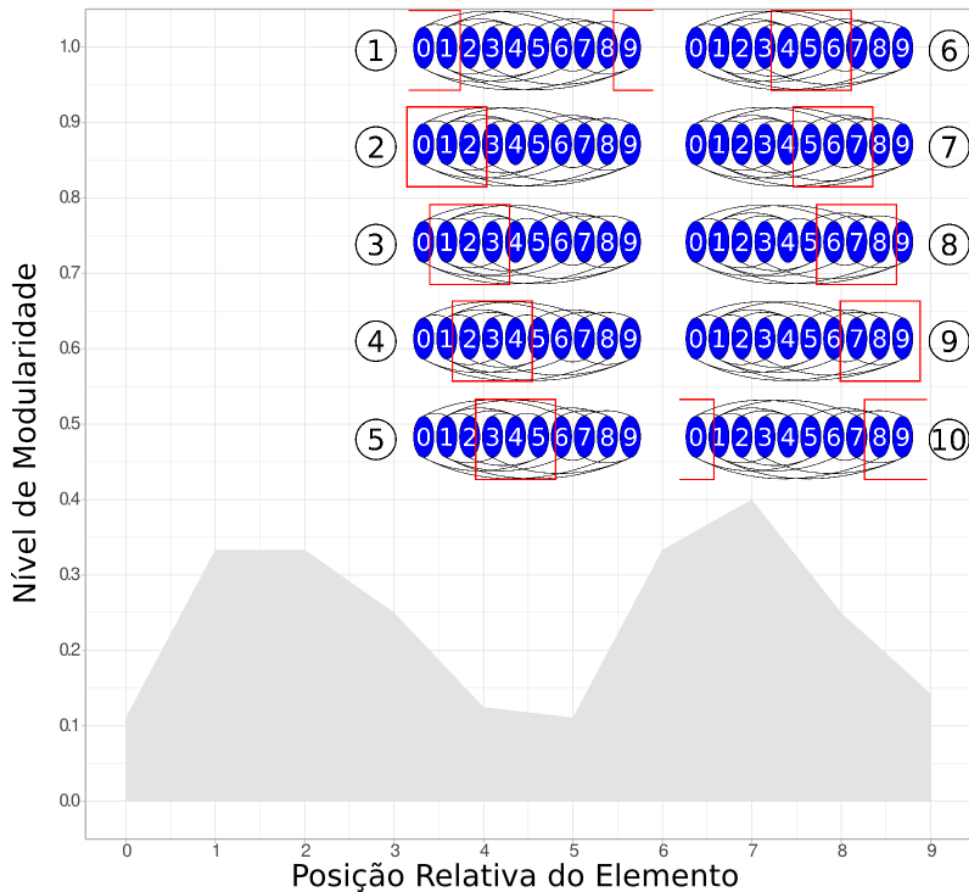


Figura 4.12: A figura mostra os agrupamentos formados pela modularidade de janela. Um primeiro agrupamento possui 0.33 de modularidade e segundo 0.4 de modularidade. Entre os dois agrupamentos temos uma máxima de 0.1 de modularidade. Isto significa que os vértices entre estes dois agrupamentos compartilham entre si apenas 10% de suas interações, lembrando que isto somente é válido para uma janela de largura $w = 3$. Fonte: (BIAZOTTI, 2016)

4.6 Análise de Expressão Gênica

Os dados de expressão gênica podem ser gerados de tecnologias de RNA-seq ou microarranjo. Após a aquisição deve-se normalizar os dados, permitindo que seja possível realizar comparações entre os resultados. Com os dados normalizados, é aplicado um filtro que remove as sondas que não estão presentes na rede proteica, caso haja referência a mais de uma sonda é aplicado um cálculo de média para a expressão das sondas. A aplicação de suavização é similar a utilizada na etapa de modularidade, distinguindo somente de não utilizar a razão no cálculo e sim a média entre os valores das expressões das sondas. A última etapa desse processo é o cálculo do p-valor que utiliza uma função beta incompleta (equação 4.8) verificando a probabilidade de ocorrência em um intervalo finito. Para o cálculo do p-valor utiliza-se os dados dos

módulos de expressão antes da suavização retornando assim o p-valor para cada sonda.

$$B_x(a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt \quad (4.8)$$

4.7 Enriquecimento Funcional

Após a aplicação da modularidade e separação dos agrupamentos, é necessário verificar a significância biológica de cada módulo. Para isto podemos utilizar bancos de dados como o Gene Ontology (GO) (ASHBURNER et al., 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (KANEHISA et al., 2012), Reactome (FABREGAT et al., 2018), entre outros. Sabemos que cada agrupamento é formado por uma coleção de vértices, os quais podem ser proteínas. Podemos utilizar ferramentas de enriquecimento funcional da plataforma **Bioconductor** (HUBER et al., 2015) e analisar cada agrupamento separadamente para encontrar os possíveis processos biológicos. Além dos processos biológicos essas ferramentas permitem a obtenção dos componentes celulares e funções moleculares.

4.8 Desenvolvimento da ferramenta *dashboard*

A ferramenta de *dashboard* visa agrupar as ferramentas do transcriptograma e do *levi* permitindo uma utilização mais intuitiva pelo usuário final.

Seu desenvolvimento foi realizado utilizando principalmente a linguagem de programação R. Outras linguagens como C e C++ foram utilizadas em algumas funções por realizarem mais rapidamente alguns processos que são custosos computacionalmente para a linguagem R, assim otimizando a velocidade de execução das tarefas. Como, por exemplo, utilizamos a linguagem C++ em algumas funções que buscam percorrer todas as células da matriz utilizando um laço de repetição.

Para facilitar o acesso a comunidade científica a ferramenta será disponibilizada no repositório do Bioconductor.

4.8.1 Método transcriptograma

O transcriptograma (RYBARCZYK-FILHO et al., 2011) é um pacote desenvolvido para o ambiente R. Esse pacote consegue analisar uma quantidade massiva de dados (*Big Data*). Foram realizadas alterações no código original do método para que os cálculos intermediários não

salvem informações temporárias necessárias para a geração do resultado final em unidades de armazenamento. Os módulos utilizados no presente trabalho são ordenamento e modularidade. Com isso as etapas realizadas pelo método são as seguintes:

1. É fornecida pelo usuário uma rede de interação, podendo ser proteína-proteína, regulação, etc;
2. O pacote do transcriptograma analisa a rede utilizando o método de Monte Carlo para encontrar *clusters* na rede em que foi fornecida pelo usuário;
3. Por fim, a ferramenta apresenta todos os *clusters* em um gráfico interativo que permite a seleção das proteínas para geração de sub-redes e apresenta métricas para análise de redes complexas.

5 *Resultados e Discussão*

5.1 *levi*

O *levi* é um pacote de código aberto, desenvolvido para o ambiente R com o propósito de visualização concomitante da projeção de expressão gênica utilizando métricas de centralidades sobre uma rede biológica. A criação do *levi* foi concebida utilizando como base dois outros *softwares*. O primeiro é o ViaComplex, um *software* que utiliza o compilador FORTRAN para criar gráficos de expressão gênica de redes (CASTRO et al., 2009). O segundo é o GALANT (GrAph LANscape VisualizaTion), um plugin para o Cytoscape que realiza as mesmas funções do ViaComplex mas com a facilidade de integração com as ferramentas já existentes do Cytoscape (CAMILO et al., 2013).

5.1.1 Implementação

Para a utilização do *levi* é necessário um arquivo contendo dados de expressão dos genes de interesse. Esses arquivos podem ser obtidos de bases de dados online (Gene Expression Omnibus (GEO), ArrayExpress, The Cancer Genome Atlas (TCGA), etc). A seleção dos genes e sua respectiva normalização dos dados de expressão devem ser realizadas pelo usuário.

O arquivo de expressão deve possuir uma coluna com dados do Gene Symbol, contendo a nomenclatura dos genes e pelo menos uma coluna com os valores de expressão. Também é possível realizar a comparação entre duas colunas de valores de expressão (Teste/Controle). Os dados dentro desse arquivo não podem estar entre aspas simples ou duplas.

Caso o arquivo de expressão não possua valores para todos os genes da rede, uma mensagem será exibida mostrando o caminho para um arquivo de log em um diretório temporário com os nomes dos genes. Na geração do landscape os genes sem valor de expressão serão exibidos com valores próximos a 0.5 demonstrando que não houveram alterações (subexpressos ou superexpressos).

Dados de rede podem ser obtidos de repositórios online (starBase, miRBase, lncRNAdb, HTRIdb, STRING, STITCH, etc). Também podem ser gerados utilizando programas específicos (Cytoscape, RedeR, Medusa, etc). O levi reconhece arquivos de rede nos formatos Medusa (DAT), RedeR (DYN), Pajek (NET) e do STRING / STITCH. Os gráficos gerados podem ser salvos nos formatos TIFF, BMP, JPEG e PNG.

O levi foi projetado levando em consideração usuários com diferentes níveis de conhecimento em informática e programação e, visando atender necessidades distintas. Por conta disso, possui dois modos de visualização: Interface de usuário (GUI) e script. O modo GUI foi implementado utilizando o pacote Shiny com o intuito de facilitar a criação de páginas web interativas e se destina principalmente a usuários com qualquer nível de conhecimento em informática. O modo script utiliza linhas de comando e permite a execução em lote para a criação de gráficos com diferentes comparações entre valores de expressão. Por utilizar o console do R, é voltado para usuários com maior experiência em programação. A versão script, por não executar o ambiente gráfico, realiza o processo de geração do gráfico um pouco mais rápido do que a versão GUI. Ambos os modos de visualização permitem alterar parâmetros relacionados ao zoom, dimensão, suavização e contraste. Os dois modos também permitem a utilização de escala logaritmica que, mantendo as proporções, melhora a visualização das expressões nas áreas da rede. Utilizando o modo de visualização GUI é possível escolher entre seis conjuntos de cores para a criação do gráfico.

5.1.2 Instalação

O pacote levi está disponível no Bioconductor através desse link. Após a instalação ele pode ser carregado utilizando o modo script ou a interface gráfica do usuário, esse último modo permite a visualização de duas formas diferentes: no navegador padrão ou com a própria engine do R. Abaixo temos as linhas de código para o carregamento e execução dos diferentes modos do pacote levi após instalação do arquivo binário.

```

1 #Instalação do levi utilizando o repositório do Bioconductor
2 if (!require("BiocManager", quietly = TRUE))
3   install.packages("BiocManager")
4
5 BiocManager::install("levi")
6
7 library(levi)
8 #Executa o levi no navegador padrao.

```

```

9 levi_ui(TRUE)
10
11 #Executa o levi no R.
12 levi_ui(FALSE)
13
14 #Executa o levi em modo script selecionando os parametros por
    linha de comando.
15 levi(expressionInput , fileTypeInput , networkNodesInput ,
        networkEdgesInput , geneSymbolInput , readExpColumn ,
        contrastValueInput , zoomValueInput , resolutionValueInput ,
        smoothValueInput )

```

5.1.3 Carregando os arquivos

O levi reconhece arquivos de rede nos formatos Medusa (DAT), RedeR (DYN), Pajek (NET) e do STRING / STITCH. O arquivo de expressão gênica deve possuir uma coluna com dados do Gene Symbol, contendo a nomenclatura dos genes e pelo menos uma coluna com os valores de expressão. Também é possível realizar a comparação entre duas colunas de valores de expressão (Teste/Controle).

A rede (nodos e interações) foi criada utilizando como base os componentes biomoleculares da via de sinalização ErbB (hsa04012) obtidos no repositório online KEGG PATHWAY que foram exportados para o STRING.

Para a representação gráfica da rede gerada, importamos os dados do STRING para o Medusa 1.5 (figura 5.1).

5.1.4 Execução do levi

Para a primeira análise realizamos duas comparações. A primeira foi com dados de expressão gênica de pessoas que não possuem adenocarcinoma mas que são fumantes em relação à pessoas que nunca fumaram e não possuem adenocarcinoma. A segunda comparação é entre pessoas que fumam e não possuem adenocarcinoma em relação à pessoas que nunca fumaram e não possuem adenocarcinoma. Para a execução do levi e geração do resultado com as comparações utilizadas nesse estudo de caso, o usuário deve digitar os seguintes comandos no console do R:

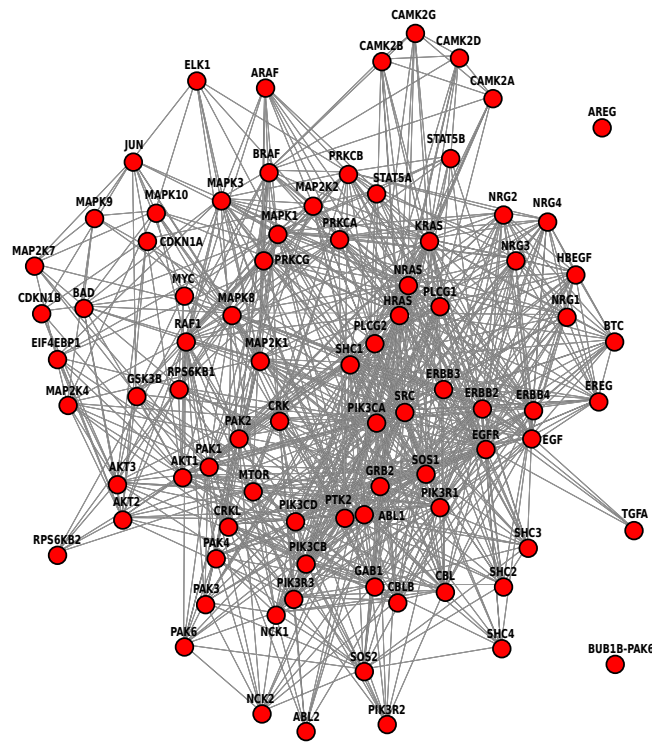


Figura 5.1: Rede gerada no STRING com componentes biomoleculares da via de sinalização ErbB e importada para o Medusa 1.5 para visualização.

```

1 library(levi)
2
3 levi(expressionInput = "GSE10072_GeneSymbol", fileTypeInput = "
  stg", networkNodesInput = "hsa04012_string_network_
  coordinates.txt", networkEdgesInput = "hsa04012_string_
  interactions.tsv", geneSymbolInput = "ID", readExpColumn =
  readExpColumn("NormalCurrentSmoker–NormalNeverSmoker", "
  NormalFormerSmoker–NormalNeverSmoker"), contourLevi = FALSE)

```

É importante observar que alguns parâmetros foram suprimidos. Se algum dos parâmetros de suavização, zoom, contraste e resolução não forem informados, eles assumem os valores padrão (valor = 50). Esse valor é utilizado por estar na metade entre os valores máximos e mínimos para cada parâmetro.

Foram realizadas três comparações para a segunda análise. A primeira foi realizada com dados de expressão de pessoas com e sem adenocarcinoma pulmonar que fumam. A segunda comparação é entre pessoas com e sem adenocarcinoma pulmonar que são ex-fumantes. A terceira comparação é entre pessoas com e sem adenocarcinoma pulmonar que nunca fumaram.

Para a execução do levi e geração do resultado com essas comparações o usuário deve digitar os seguintes comandos no console do R:

```
1 library(levi)
2
3 levi(expressionInput = "GSE10072_GeneSymbol", fileTypeInput = "
  stg", networkNodesInput = "hsa04012_string_network_
  coordinates.txt", networkEdgesInput = "hsa04012_string_
  interactions.tsv", geneSymbolInput = "ID", readExpColumn =
  readExpColumn("TumorCurrentSmoker-NormalCurrentSmoker", "
  TumorFormerSmoker-NormalFormerSmoker", "TumorNeverSmoker-
  NormalNeverSmoker"), contourLevi = FALSE)
```

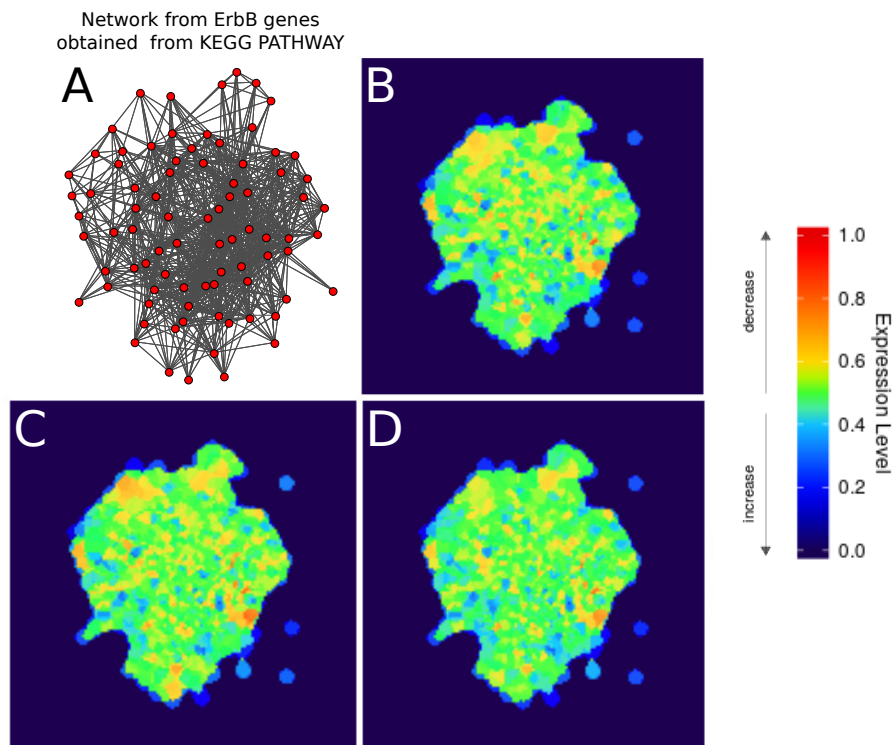


Figura 5.2: Utilizando como entrada os arquivos de coordenadas e interações obtidos do STRING e os dados de valores de expressão de adenocarcinoma pulmonar o levi gerou como resultado uma visualização integrada dos dados em formato landscape. Em (A) temos a visualização dos nós e interações da rede pelo software Medusa. Temos em (B) os resultado do levi para as condições "fumante com câncer em relação a indivíduos sem câncer e que nunca fumaram", (B) "ex-fumante com câncer em relação a indivíduos sem câncer e que nunca fumaram" e (C) "indivíduos que nunca fumaram com câncer em relação a indivíduos sem câncer e que nunca fumaram".

O levi é uma aplicação desenvolvida em R, a qual pode ser executada em qualquer computador com Windows, Linux ou Mac que possuam esse ambiente instalado. Por possuir dois modos de execução (script e interface gráfica) atende as necessidades de usuários com diferentes níveis de experiência em programação. A visualização integrada resultante do processamento do levi entre os dados de expressão e a rede possibilitam uma análise mais completa da interação dos genes, além de oferecer a opção de integrar o levi com outros pacotes disponíveis para o R, aumentando o leque de possibilidades e análises que a aplicação disponibiliza.

5.1.5 Bioconductor

O levi está disponível para *download* no Bioconductor. O Bioconductor (<https://bioconductor.org/>) é um repositório *online* de pacotes de bioinformática com código fonte aberto para serem utilizados em R. O levi foi aceito e publicado em novembro de 2018 e pode ser acessado através do link: <https://bioconductor.org/packages/levi/>.

A tabela 5.1 demonstra a quantidade de *downloads* do pacote levi no Bioconductor no período de novembro de 2018, data de sua inclusão no repositório, até agosto de 2022.

Tabela 5.1: Número de *downloads* por ano e mês do pacote levi no Bioconductor.

Ano	Mês	Computadores diferentes	Total de <i>downloads</i>
2018	novembro à dezembro	52	124
2019	janeiro à dezembro	301	847
2020	janeiro à dezembro	339	1114
2021	janeiro à dezembro	456	1097
2022	janeiro à agosto	195	434
Total	-	1346	3616

Fonte: O autor (2022).

5.2 transcriptograma

Foi necessário fazer alterações visando a adequação para o aceite do pacote no repositório Bioconductor. Dois dos pontos principais de alterações foram o armazenamento resultante para os processos de ordenamento e modularidade. Os resultados desses processos foram inseridos em um objeto do R. Esse objeto é uma lista de listas com diferentes estruturas resultantes dos processamentos realizados.

Na figura 5.3 podemos observar a estrutura do objeto criado para armazenar os resultados do ordenamento (*ordering_data*). Para cada semente de número aleatório (*seed*) é gerado uma

lista, dentro dessa lista são geradas novas listas com os dados da rede, ordenamento, tempo de execução, energia, temperatura, trocas, fator de resfriamento, intervalo de resfriamento e percentual de energia.

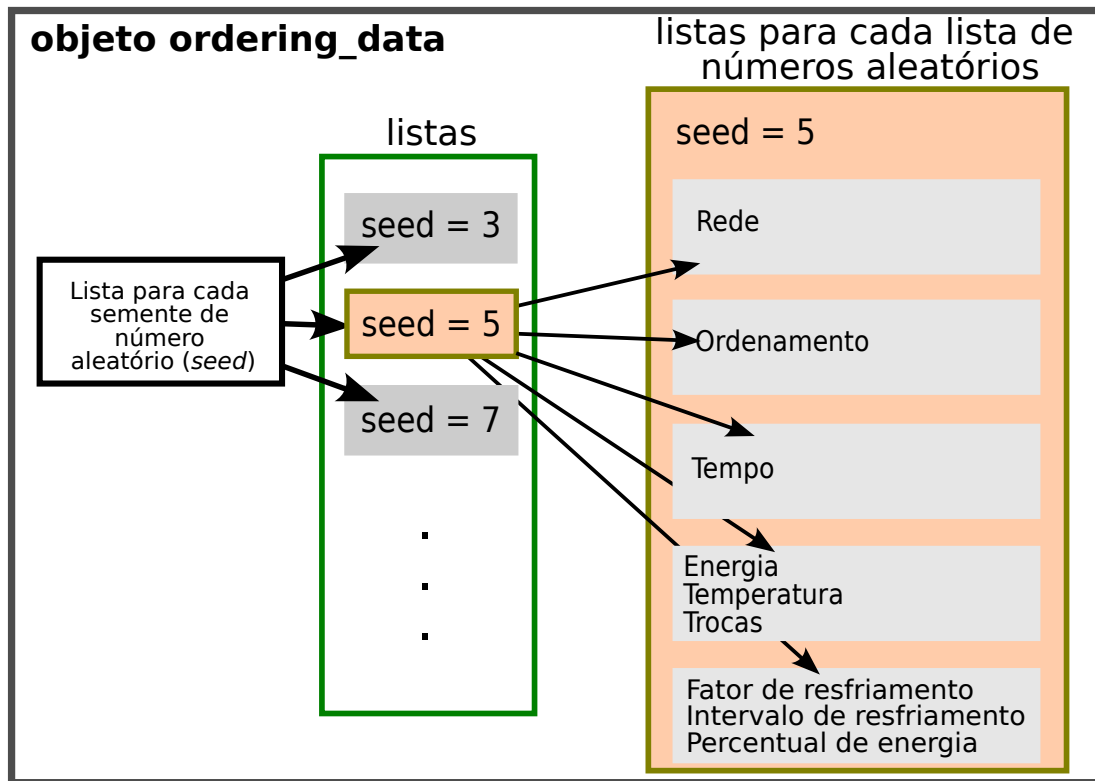


Figura 5.3: Objeto *Ordering Data*. Objeto do R contendo listas de listas para armazenamento dos resultados do ordenamento com os dados da rede original, resultado do ordenamento, tempo de execução, energia final, temperaturas, trocas realizadas, fator de resfriamento, intervalo de resfriamento e percentual de energia.

Após a etapa de execução do ordenamento e com o objeto criado, é possível realizar o processo de modularidade. Os resultados do processamento de modularidade também são armazenados em um outro objeto de R que é formado por listas dentro de listas. Para cada janela é gerado uma lista que possui um campo de identificador único para cada elemento e o valor de modularidade daquele elemento (5.4)

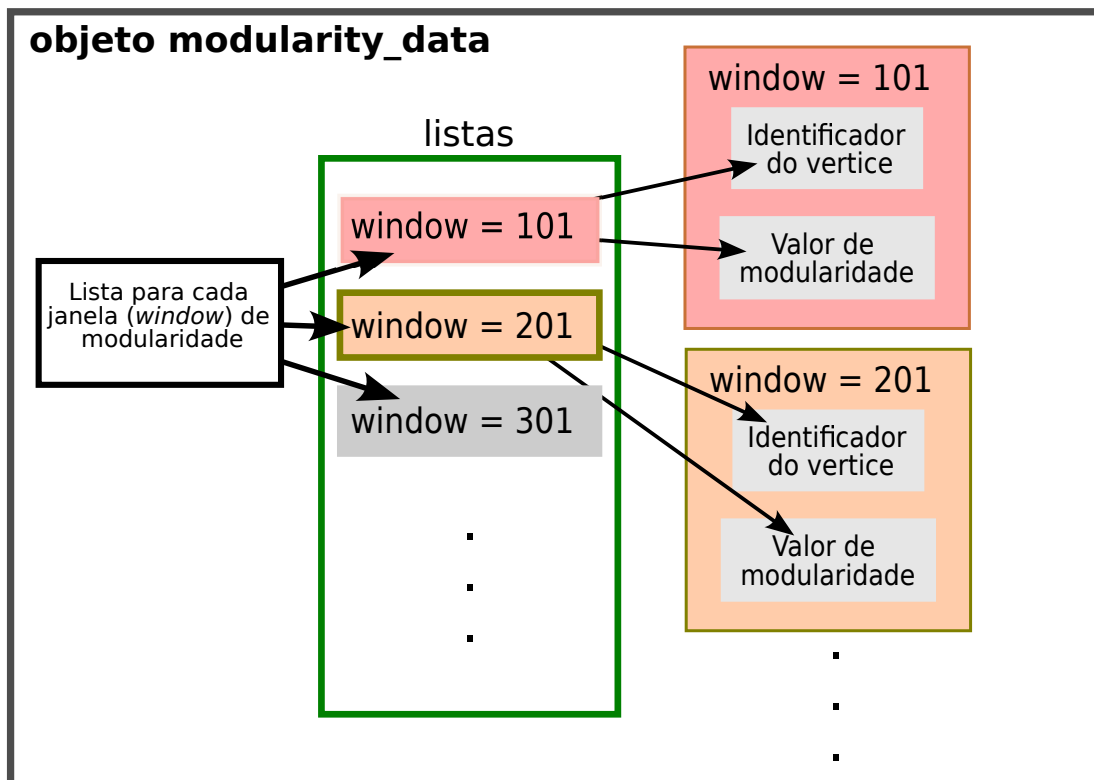


Figura 5.4: Objeto *Modularity Data*. Objeto do R contendo listas de listas para armazenamento dos resultados da modularidade. Para cada valor de janela selecionada são armazenados o identificador do vértice e o valor de modularidade.

5.3 Ferramenta de *dashboard*

A ferramenta de *dashboard* visa facilitar a utilização, integração e visualização dos resultados das ferramentas *levi* e *transcriptograma*. O *workflow* para execução do dashboard possui as seguintes etapas:

- Execução da função de ordenamento no arquivo contendo os dados de interação da rede.
- Utilização da função de modularidade para gerar as janelas com os tamanhos de interesse.
- Com os resultados de ordenamento e modularidade o *dashboard* poderá ser iniciado.

Abaixo temos os comandos exemplificando as referentes etapas para execução do *dashboard*:

```
1 library(rTranscriptogram)
```

```
2
```

```

3 ordering <- ordering_rtrans(network_data = "yeast_08.txt",
   model = "c" , mcstep = 100, seed = c(3, 5, 7), log = 10)
4 modularity <- modularity_rtrans(network_data='yeast_08.txt',
   window = c(101, 151, 201,251), ordering_result = ordering)
5
6 module_rtrans(ordering , modularity)

```

As funções de ordenamento e modularidade necessitam receber alguns parâmetros para execução. Para a função de ordenamento é necessário informar o nome do arquivo contendo os dados de interação da rede (*network_data*). O modelo de ordenamento (*model*), que pode ser dos tipos: Anel (a), Cruz (c) e X (x). Quantidade de passos de Monte Carlo (*mcstep*). A semente para os números aleatórios (*seed*). E o log que equivale a quantidade de execuções.

A função de modularidade também recebe os parametro de nome do arquivo contendo os dados de interação da rede (*network_data*). A dimensão da janela é informada no parâmetro *window*, sendo que quanto maior for a janela menor a quantidade de picos de energia no transcriptograma. E é necessário informar a variável onde estão os resultados do ordenamento (*ordering_result*).

A tela do *dashboard* pode ser executada de forma nativa usando o próprio R ou pode ser executada no navegador de internet padrão. A escolha é realizada no momento da inicialização do *dashboard* usando o parâmetro “F” (*false*) para utilizar a visualização com o próprio R ou “T” (*true*) alterando assim para o navegador de internet padrão. Ao final será exibida a tela conforme figura 5.5.

Na figura, temos as seguintes sessões separadas para melhor identificação: (A) menu lateral de navegação entre as opções integradas do *dashboard* contendo: *Ordering and Modularity*, *Dashboard*, *Cluster Plot*, *Gene Expression Level*, *Functional Enrichment* e o programa *levi*. (B) seleção da janela de modularidade processada em etapa prévia. (C) gráfico dinâmico com a visualização dos níveis de modularidade em relação a posição relativa das proteínas da rede, no qual é possível selecionar conjuntos de proteínas nesse gráfico ou inserir valor inicial e final para cada *cluster*, nomeá-los e criar um *cluster* para análise posterior. (D) Área de exibição e gerenciamento dos *clusters*. Apresenta o nome dos *clusters*, posição inicial e final da proteína, onde também é possível apagar *clusters* utilizando o sinal de “traço” existente na coluna *delete*. (E) Tabela contendo os nomes das proteínas selecionadas no gráfico dinâmico, sua posição relativa e nível de modularidade. Logo acima da tabela existem botões para busca e exportação desses dados, é possível copiar, imprimir ou exportar para os formatos csv (*comma separated*

values) e pdf.

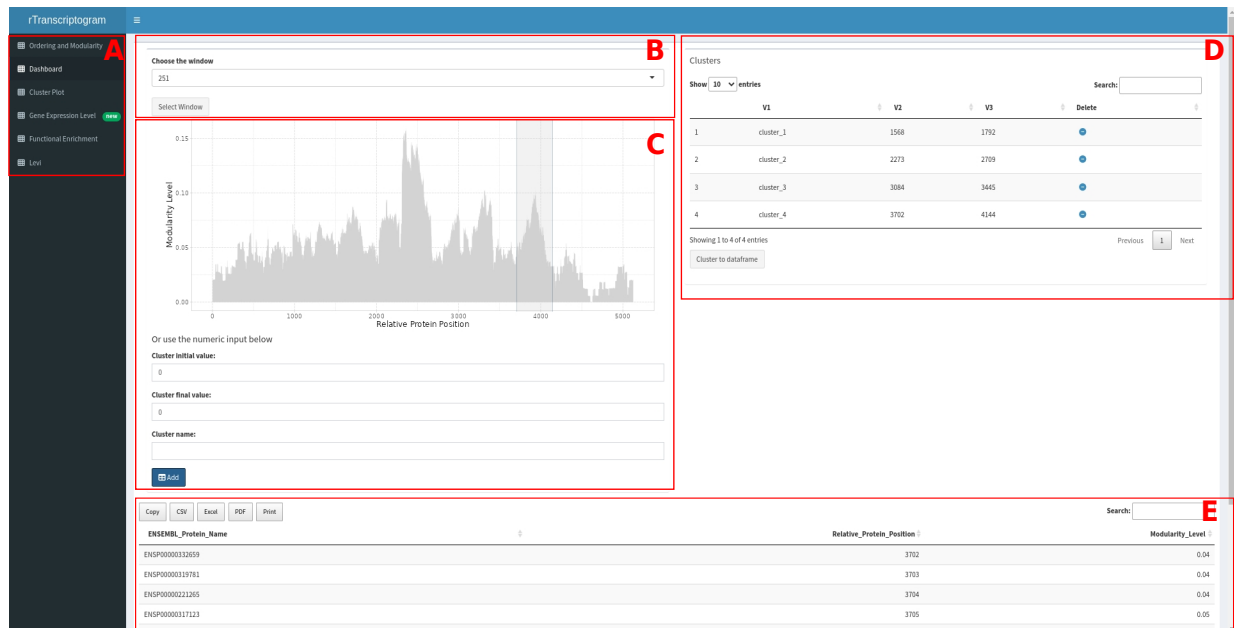


Figura 5.5: Tela principal de execução do *dashboard*. A opção de *dashboard* (A) possui a interface gráfica para a seleção da janela (B), visualização da posição relativa da proteína e seus respectivos níveis de modularidade (C), agrupamentos selecionados (D) e uma tabela detalhada com o nome da proteína, posição relativa e nível de modularidade (E).

Após a seleção dos *clusters*, o botão *Cluster to dataframe* é habilitado. Esse botão cria uma variável no R do tipo *dataframe* que armazena os dados dos *clusters* gerados, mesmo após finalizada a execução do programa essa variável fica disponível para o usuário podendo realizar outras análises. Após gerar o *dataframe* é possível selecionar a opção *Cluster plot* e *Gene Expression Level*, localizados no menu lateral. As opções existentes nessas telas só estão disponíveis após a geração do *dataframe* com os *clusters*.

A figura 5.6 exemplifica a execução da tela de *Cluster plot*, a qual está dividida nas seguintes seções: (A) área de seleção dos *clusters* criados através dos cortes de seleção no gráfico dinâmico na opção *dashboard*. Quando é selecionado os dos *clusters* existentes são apresentadas as métricas para a geração da rede de interação entre as proteínas: quantidade de vértices, quantidade de arestas, coeficiente de clusterização global, coeficiente de clusterização médio, diâmetro da rede, raio da rede, centralização da rede, caminho mais curto entre os vértices, densidade da rede e se existem vertices isolados. (B) gráfico dinâmico com a interação entre as proteínas. (C) tabela com os nomes e as interações entre as proteínas do *cluster* que formam a rede exibida no gráfico dinâmico. Logo acima da tabela também existem botões para busca e exportação desses dados, é possível copiar, imprimir ou exportar para os formatos csv (*comma separated values*) e pdf.

Ao ser selecionado os *clusters* na opção do menu do *Dashboard* as opções de entrada do menu *Gene Expression Level* são habilitadas. A opção apresentada na figura (figura 5.8) (A), *Upload the treated files*, permite ao usuário carregar os dados já suavizados e com os nomes das proteínas do experimento já modificados para serem referentes a rede de proteínas utilizada no ordenamento. O padrão de arquivo aceito para essa opção é o csv (*comma separated values*).

Caso o usuário não possua o arquivo pronto para ser carregado na ferramenta é possível utilizar a opção existente na figura 5.8 (B), *Upload separated files*, onde arquivos de *microarray* ou RNA-seq podem ser utilizados. Para o *microarray*, é possível selecionar vários arquivos contendo os dados de experimento. Para o RNA-seq é necessário que seja apenas um arquivo com os filtros dos atributos de interesse. Os dados carregados dos arquivos de experimento são apresentados na figura 5.8 (C).

The screenshot shows the rTranscriptogram web application. On the left is a sidebar with navigation options: Ordering and Modularity, Dashboard, Cluster Plot, Gene Expression Level (highlighted with a 'new' badge), Functional Enrichment, and Level. The main content area has three tabs: Files, Biomart and Suavization, Plot Settings, and Plot. The 'Files' tab is active, showing two upload options: 'Upload the treated files' (A) and 'Upload separated files' (B). Below these is a table (C) displaying gene expression data.

gene_id	L400T	L401T	L404T	L406T	L413T	L414T	L417T	L420T	L439T	L440T	L441T	L442T	L444T	L446T	L447T	L452T	L455T	L456T
ENSG000000000003	52.195	37.8891	23.191	25.0324	41.9686	28.5794	19.9008	23.2632	37.024	101.314	15.6387	37.5692	30.9044	32.6581	43.0741	19.067	44.9344	12.8815
ENSG000000000005	0.230061	0.0860342	0.0480224	0	2.5709	0.0871915	0.234047	0	0	0	0	0.0880154	0.0945076	0.0323317	0.341157	0	0	0
ENSG0000000000419	43.8616	47.0457	38.1292	54.303	51.2969	42.1604	78.1961	60.7283	23.5296	34.258	43.5934	51.5559	42.0803	55.831	83.7509	79.8063	33.4649	44.7526

Figura 5.8: Tela inicial da opção do menu *Gene expression Level*. Permite o carregamento de arquivos de microarranjo ou RNA-seq já tratados com a aplicação da suavização e padronização dos nomes em referência aos dados da rede de proteínas (A) ou arquivos brutos ainda não tratados (B), caso sejam selecionados os arquivos com os dados brutos é possível ver um resumo dos dados carregados (C).

A aba *Biomart and Suavization* do menu *Gene Expression Level* utiliza o pacote Biomart do repositório Bioconductor que busca facilitar a mineiração de dados entre diversos tipos de conjuntos de dados disponíveis na *internet* permitindo assim a anotação de genes e pesquisar entre os nomes das tecnologias existentes aqueles que são sinônimos para um resultado padronizado.

Nos campos existentes na figura 5.9 (A), podemos selecionar os repositórios *online* em *List Marts* e as opções de bancos de dados disponíveis (*Select Data Set*) atualiza automaticamente de acordo com a seleção do repositório. Entre os repositórios disponíveis temos:

- **ENSEMBL MART ENSEMBL** - possui dados para anotação geral de diversos organismos, como por exemplo o *Homo sapiens*;
- **ENSEMBL MART MOUSE** - possui dados para anotação funcional de genomas de *Mus Musculus*;
- **ENSEMBL MART SNP** - possui dados para anotação funcional de polimorfismos de nucleotídeos;
- **ENSEMBL MART FUNCGEN** - possui dados para anotação funcional de regiões reguladoras ou relações de genes.

O biomart precisa de uma conexão com a *internet* para acessar as bases de dados dos seus repositórios, caso a conexão com o servidor principal esteja apresentando lentidão o próprio pacote busca se conectar a servidores secundários sem a necessidade de interação com o usuário.

Na figura 5.9 (B), podemos visualizar os campos referentes ao filtro e atributos que serão utilizados no processamento do Biomart. O campo filtro possui a função de retornar os resultados de similaridades dos bancos de dados selecionado utilizando o atributo escolhido nessa etapa da consulta. A tabela apresentada na opção *Select attributes by clicking on the lines* deve possuir novamente o filtro selecionado na etapa anterior e o identificador que será utilizado para a conversão dos nomes que é o mesmo padrão utilizado na rede biológica.

Para a suavização, utiliza-se o conceito de janela apresentado na etapa da modularidade (figura 4.11). A diferença desse processo em relação a suavização de dados na modularidade é a utilização da média onde são somados os valores de expressão existentes em cada sonda e divididos pelo mesmo valor de janela utilizado na etapa da modularidade.

Após ser recuperado os dados das bases de dados utilizando o biomart, realizado a referência dos nomes das sondas existentes nos arquivos de dados de expressão com os nomes na rede biológica e aplicado a suavização a aba *Plot Settings* é habilitada (figura 5.10). Essa aba possui os parâmetros para geração do gráfico de expressão suavizado. É necessário selecionar na tabela *Table and Plot* as colunas referentes aos dados iniciais de cada experimento, a quantidade de colunas selecionadas deve estar relacionada ao número de comparações. É possível adicionar uma linha de referência no gráfico utilizando operadores relacionais em comparação a quantidade de vezes o desvio padrão selecionado. Para finalizar, a opção de número de comparações permite selecionar o tamanho de cada segmento que será utilizado para a geração do gráfico.

A última aba do menu *Gene Expression Level* possui os gráficos gerados através do processamento dos dados dos arquivos de expressão (figura 5.11). Na mesma tela, logo abaixo da

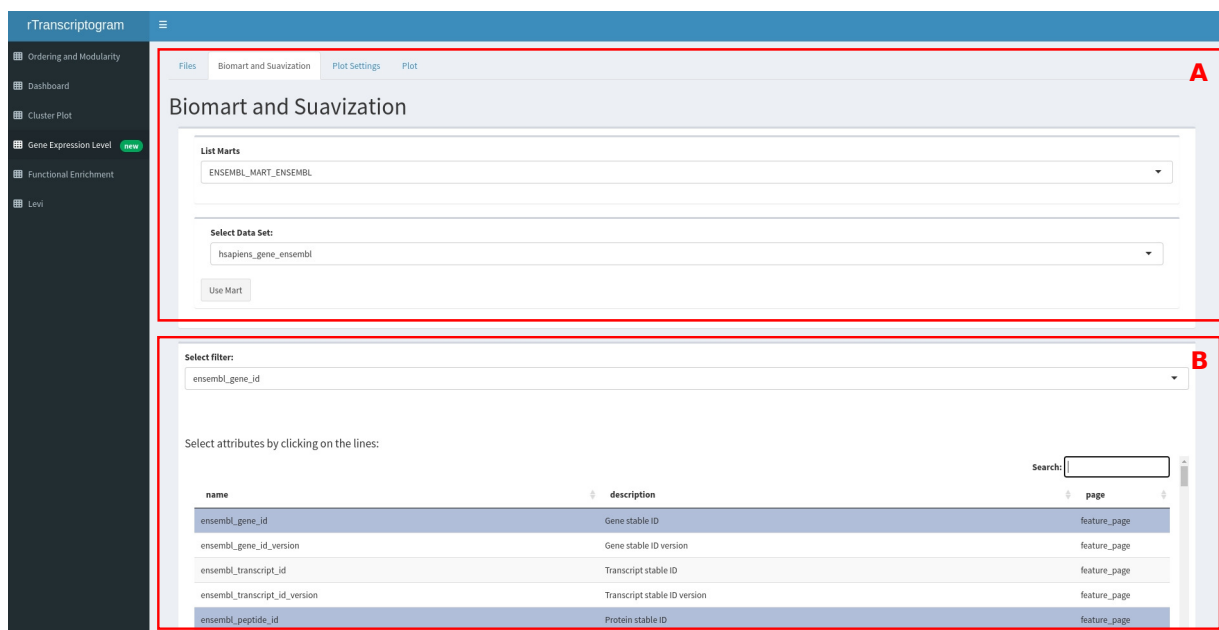


Figura 5.9: Aba *Biomart and Suavization* da opção do menu *Gene Expression Level*. Utiliza o pacote Biomart do Bioconductor para padronização dos nomes da rede biológica com os dados de expressão. As etapas são realizadas em sequência, primeiro é necessário selecionar o *mart*, assim são carregados as bases de dados disponíveis (A). Após carregado as bases de dados é possível selecionar os filtros e atributos de interesse (B).

figura com todos os gráficos gerados pelos experimentos, é possível selecionar separadamente cada gráfico aumentando assim a visualização e facilitando as análises (figura 5.12).

O menu *Functional Enrichment* é habilitado após a geração dos gráficos de nível de expressão gênica. Na primeira aba, *Targets*, é possível a seleção da quantidade de comparações a serem realizadas, assim como nomear as comparações (nomes são gerados automaticamente caso o usuário não deseje fazer alterações) e definir para cada amostra os valores de referência e teste (figura 5.13).

Na figura 5.14 é apresentado a tela de seleção de parâmetros para a ontologia e e processamento estatístico. É possível selecionar entre três tipos de ontologia:

- **Cellular Component** - Localização relativa às estruturas celulares onde os produtos de genes desempenham alguma função como por exemplo a mitocôndria ou o ribossomo.
- **Biological Process** - Processos realizados por múltiplas atividades moleculares como por exemplo o reparo de DNA.
- **Molecular Function** - Atividades realizadas por produtos de genes em nível molecular como transporte ou catálise.

Também na aba *Ontology and Statistic* deve ser selecionado a base de dados para que seja

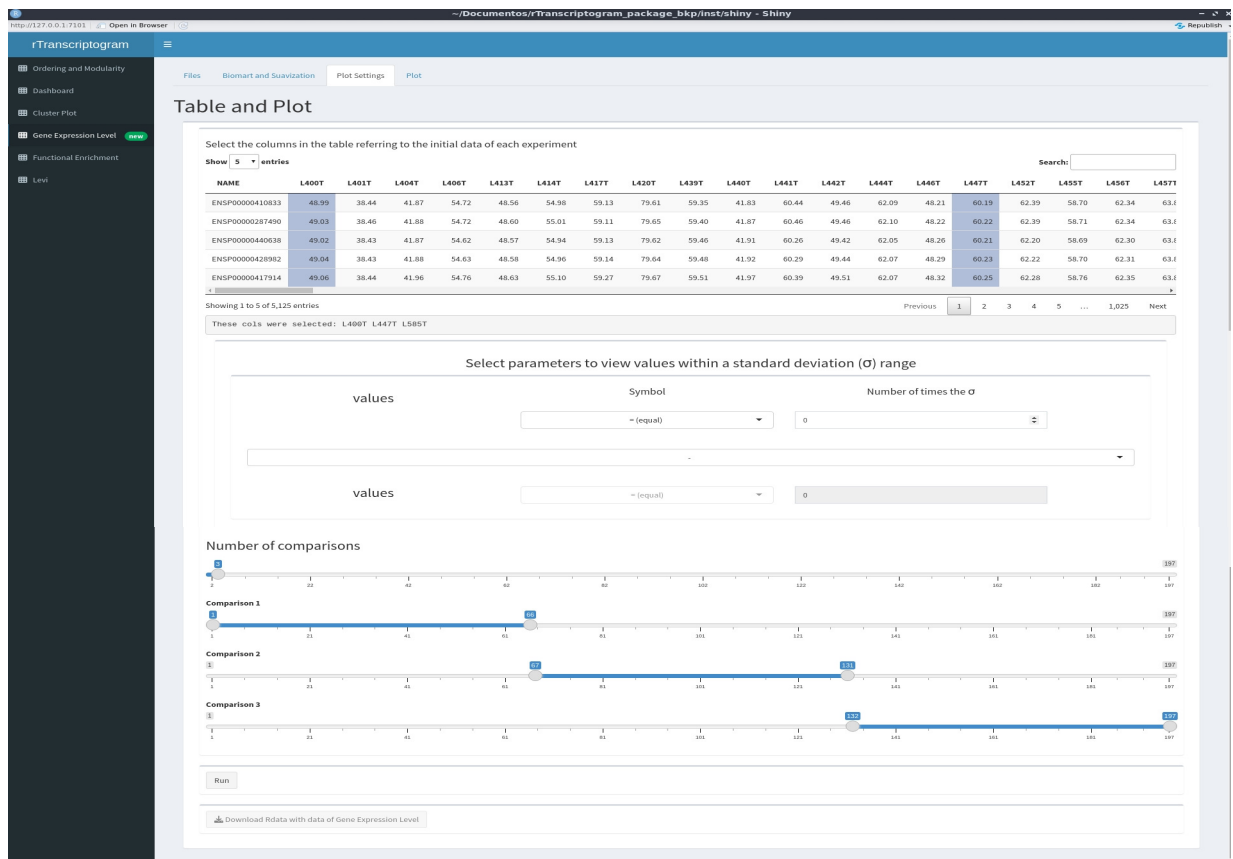


Figura 5.10: Aba *Plot Settings*. Possui a interface de parametrização para geração do gráfico de expressão. Nela é possível visualizar o resultado da aplicação da suavização e troca de nomes das sondas, caso tenha sido realizado essa etapa, assim como aplicar linhas de referência em relação aos valores de desvio padrão e a quantidade de comparações existentes para os dados de expressão.

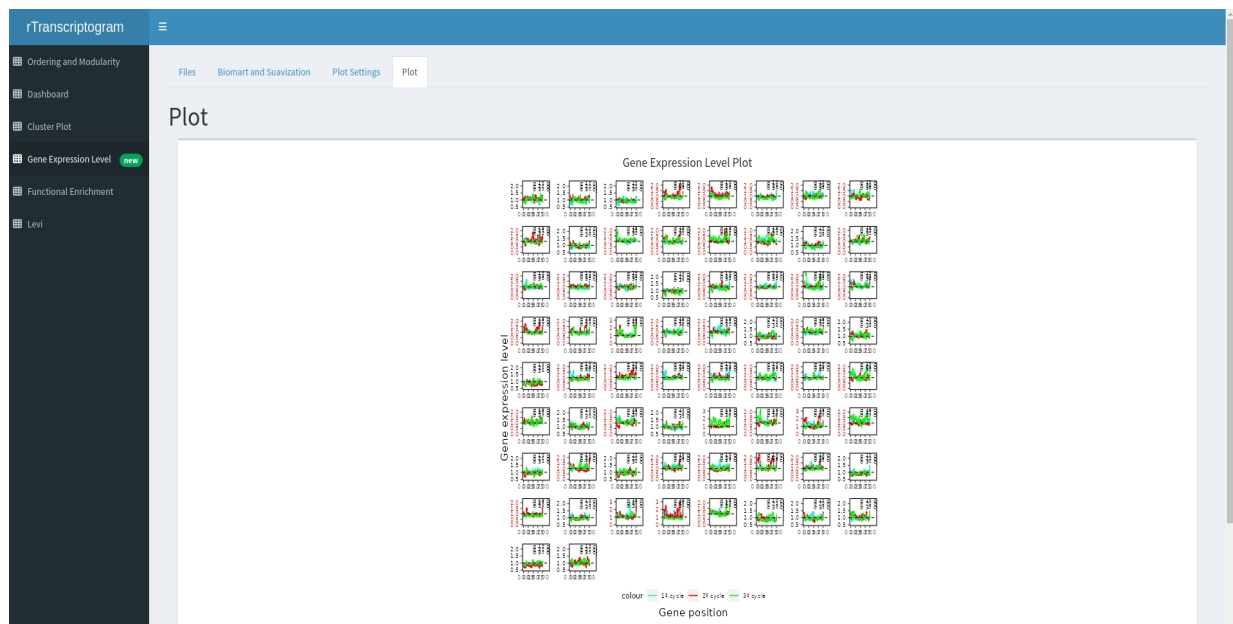


Figura 5.11: Agrupamento do resultado da geração dos gráficos de expressão gênica.

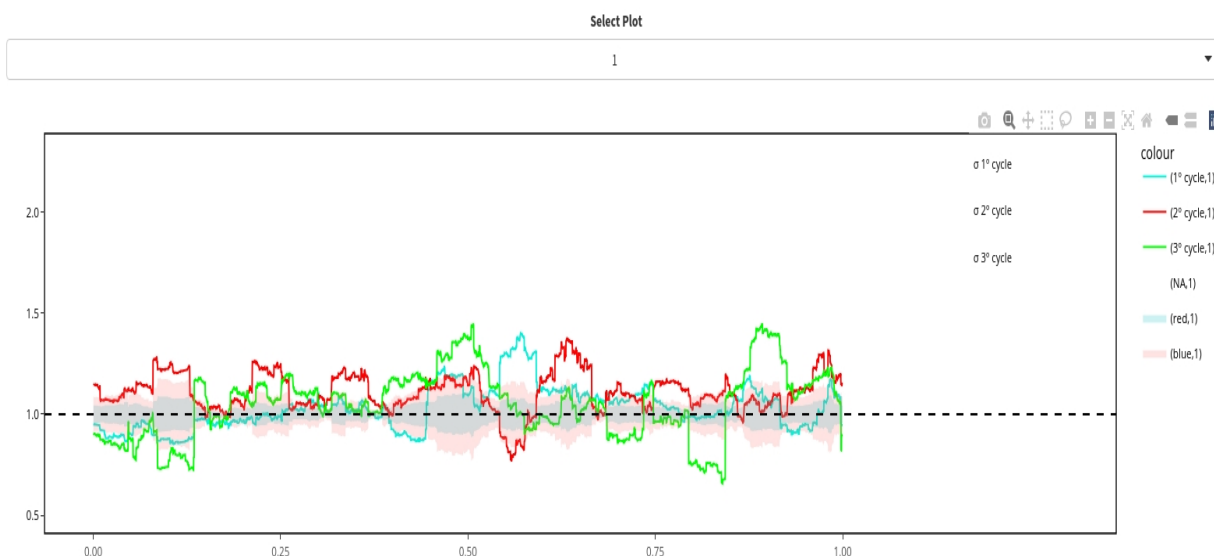


Figura 5.12: Seleção unitária dos gráficos de expressão gênica.

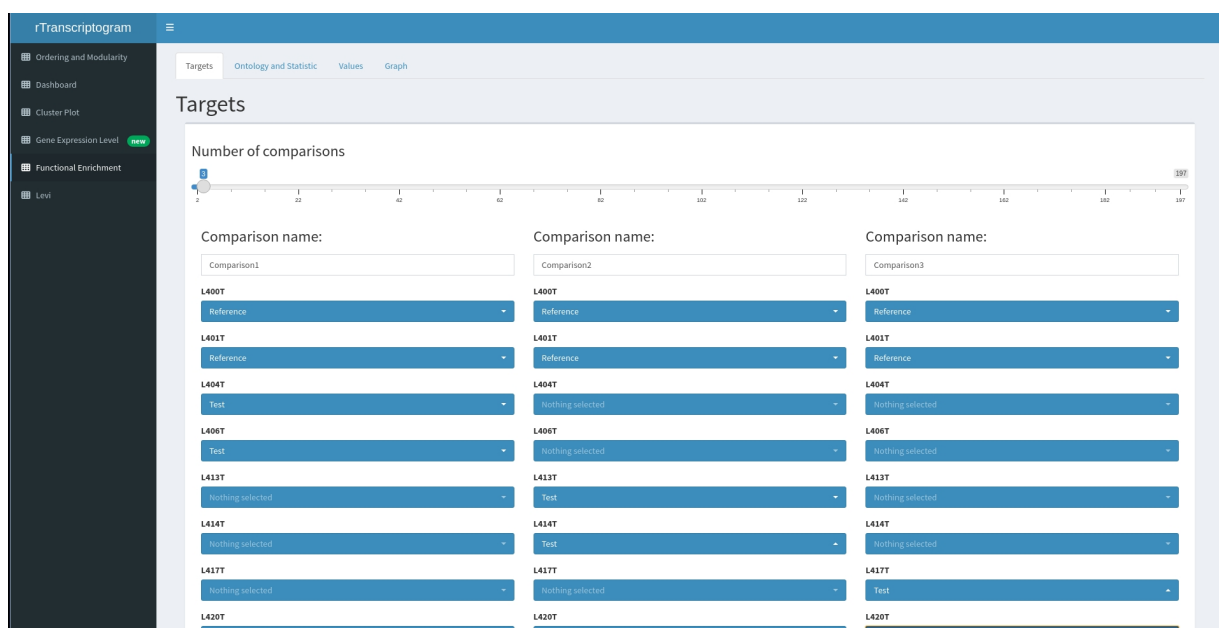


Figura 5.13: Aba *Targets* do menu *Functional Enrichment*. Através da seleção de quantidade de comparações são gerados automaticamente os campos com os nomes dos arquivos que podem ser relacionados como sendo de referência ou teste. Também é possível alterar o nome de cada comparação na caixa de edição localizada acima de cada coluna de comparação.

realizado o mapeamento entre os dados de expressão e a ontologia. O parâmetro *Ratio Cutoff* permite definir um valor para que seja realizada a filtragem de genes com significância estatística mas que estejam abaixo de um determinado valor de *fold-change*. Para finalizar é necessário escolher o p-valor para a filtragem da seleção das proteínas.

A aba *Comparison Results* (figura 5.15) possui os resultados do processamento realizado

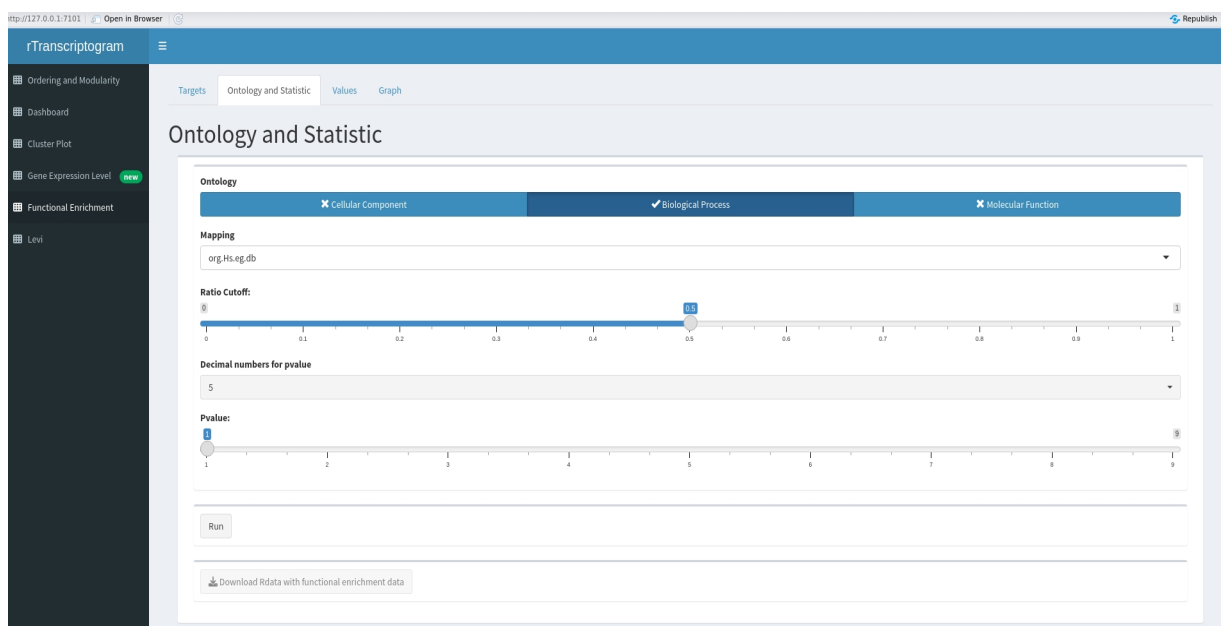


Figura 5.14: Aba *Ontology and Statistic* do menu *Functional Enrichment*. Permite a seleção do tipo de ontologia entre componentes celulares, processos biológicos e função molecular. A base de dados para mapeamento dos dados de expressão. O valor da razão entre a quantidade de proteínas em relação ao número total de proteínas presentes na ontologia. O p-valor para filtragem dos resultados.

para as ontologias e filtrações usando o p-valor. É possível selecionar qual resultado de comparação deseja visualizar através da opção *choose the comparison*. Ao selecionar a comparação é exibido uma listagem com todas as GO's permitindo selecionar ou excluir as ontologias antes da geração do gráfico final.

A última aba do menu *Functional Enrichment*, chamada *Graph* (figura 5.16), possui a integração dos resultados obtidos da seleção dos cluster com os dados de expressão. Temos no gráfico superior a atividade transcricional relativa e no gráfico inferior os p-valores. Em ambos os gráficos, podemos verificar as posições das GO's selecionadas na aba *Results*.

A opção no menu lateral *levi* já é habilitada a partir da execução do rTranscriptograma, por ele ser um módulo que pode ser executado independente das etapas anteriores. Contudo, para utilizar a rede gerada pelas etapas de ordenamento e modularidade é necessário a criação de pelo menos um cluster e a seleção do mesmo na aba *Cluster Plots*. Com os dados de interação de proteínas transformados em uma rede pode-se utilizar a mesma rede como entrada para o *levi* selecionando o parâmetro "Transcriptogram" em "Network input type". Com os dados da rede carregados é necessário informar apenas o arquivo contendo os valores de expressões para a geração do resultado da rede com a sobreposição dos valores de expressão de teste e/ou controle. Como é possível observar na figura 5.17 o layout da rede é o mesmo visto na figura 5.6 (B).

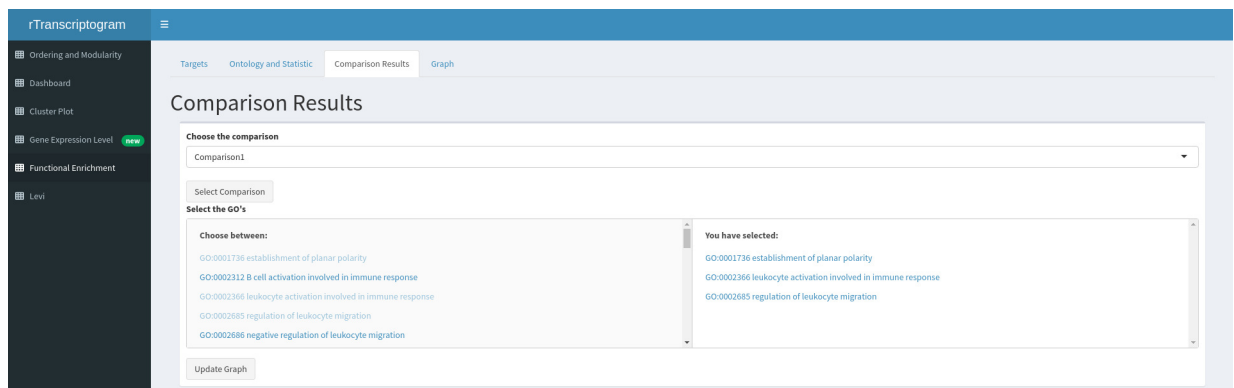


Figura 5.15: Aba *Results* do menu *Functional Enrichment*. Permite a seleção das comparações. Para cada comparação selecionada apresenta todas as ontologias encontradas. É possível selecionar as ontologias escolhidas para serem visualizadas no gráfico do rTranscriptograma.



Figura 5.16: Aba *Graph* do menu *Functional Enrichment*. Apresenta o resultado final do rTranscriptograma. Acima possuímos um gráfico com a atividade transcritora relativa em relação a posição relativa de cada elemento. No eixo y temos uma linha de referência no valor um, valores acima dessa linha de referência estão superexpressos e valores abaixo dessa linha de referência estão subexpressos. No gráfico abaixo temos os p-valores para cada posição relativa dos elementos. Em ambos os gráficos são exibidos segmentos de setas demonstrando a posição das ontologias.

5.4 Aplicação do *workflow* em dados de *Homo sapiens*

A rede para *Homo sapiens* obtida no STRING possui como características 11.468 proteínas que formam um total de 277.015 ligações entre elas. Utilizando essa rede aplicamos as funções

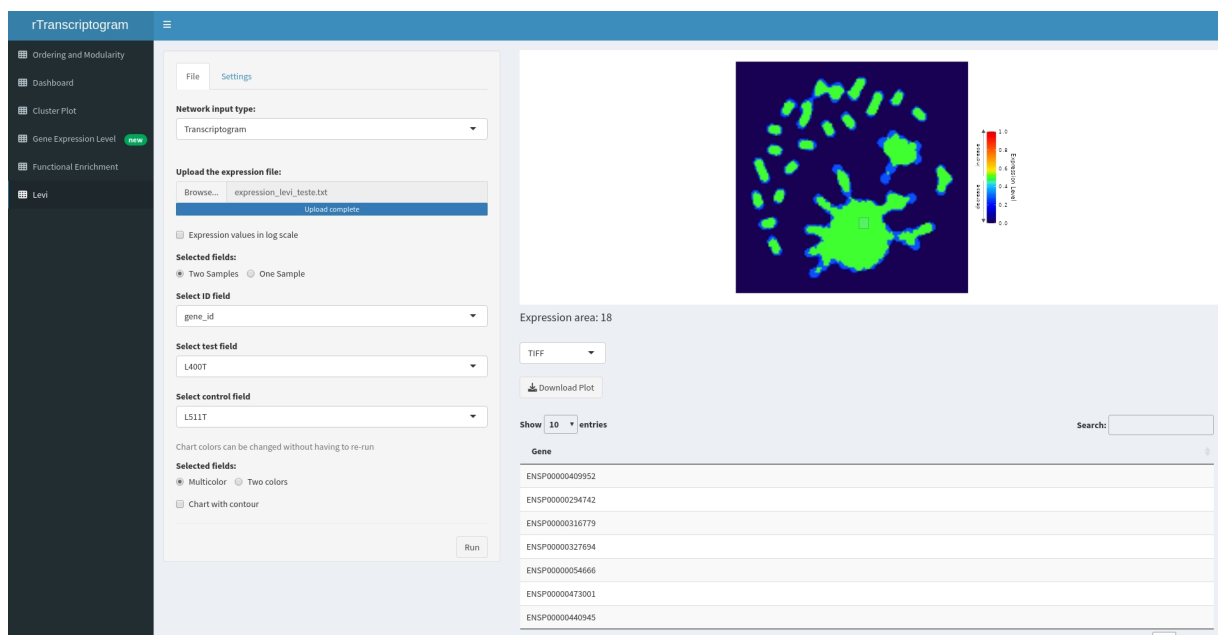


Figura 5.17: Exemplo de visualização do levi dentro do *dashboard* utilizando como base a rede de interação de proteínas criada pelo rTranscriptograma. No levi podemos selecionar o tipo de rede entre transcriptograma, medusa, redeR, Pajek e String/Stich. Selecionar o arquivo com os dados de expressão. Quantidade de amostras. Campos referentes ao identificador, teste e controle. Tipos de conjuntos de cores e contorno na geração do gráfico.

de ordenamento e modularidade. A função de ordenamento foi realizada usando o modelo em Cruz, Anel e X com 10.000 passos de Monte Carlo, semente para os números aleatórios (*seed*) nos valores de 251 além de log no valor 10. Para a modularidade usamos as janela com o valores de 101, 201, 301, 401 e 501.

As figuras 5.18, 5.19 e 5.20 apresentam as modularidades para, respectivamente, os modelos Cruz, Anel e X utilizando as janelas 101 (em tom de cor vermelho), 201 (em mostarda), 301 (em verde), 401 (em azul) e 501 (em violeta). Podemos verificar que existem similaridades na formação de agrupamentos e picos em todas as janelas de cada um dos modelos. Contudo, a janela de tamanho 101 resultou na formação de agrupamentos com picos com maior facilidade de identificação.

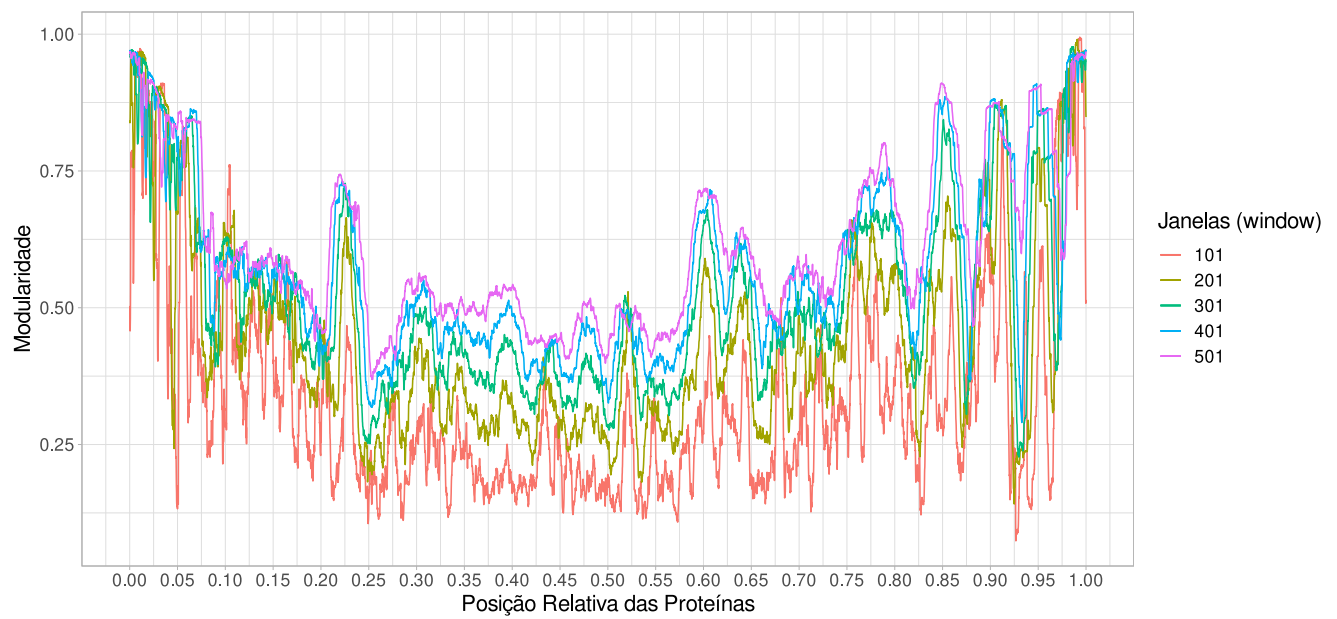


Figura 5.18: Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo Cruz.

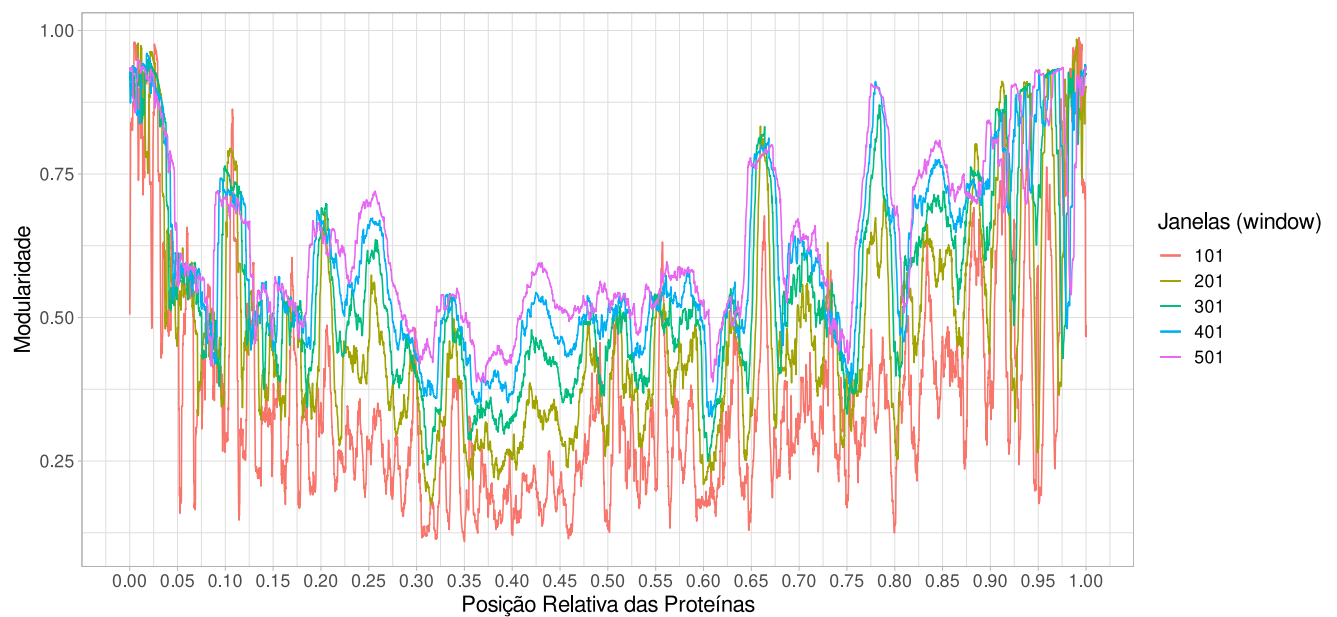


Figura 5.19: Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo Anel.

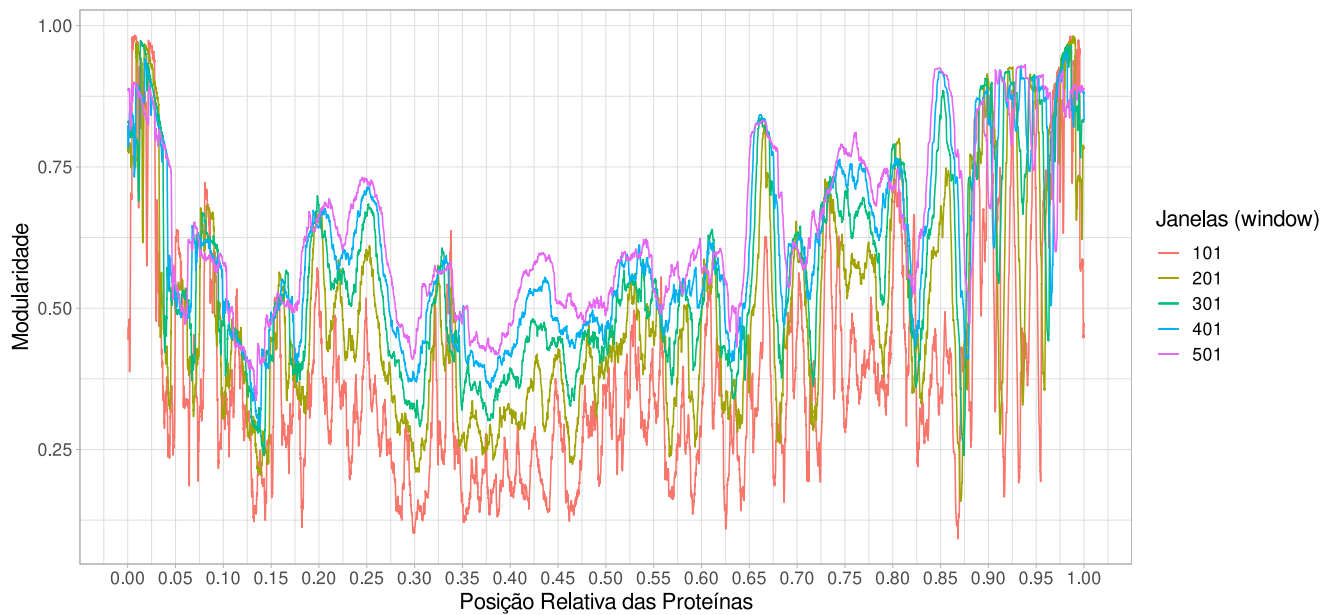


Figura 5.20: Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo X.

A figura 5.21 apresenta o custo energético em relação a quantidade de passos de Monte Carlo realizados. As linhas representam os modelos utilizados (vermelho para o modelo Cruz, verde para o modelo Anel e azul para o modelo X). O eixo do custo energético está em log de base 10. É possível verificar que o modelo Anel, que faz a verificação dos vizinhos nas oito possíveis direções (figura 4.10), possui o maior custo energético em relação aos três modelos aplicados a rede *Homo sapiens*, seguido pela modelo X que verifica os vizinhos das diagonais superiores e inferiores (figura 4.9) e o modelo Cruz, que verifica os vizinhos superior, inferior, esquerdo e direito em relação ao nodo central (figura 4.7), com o menor custo energético.

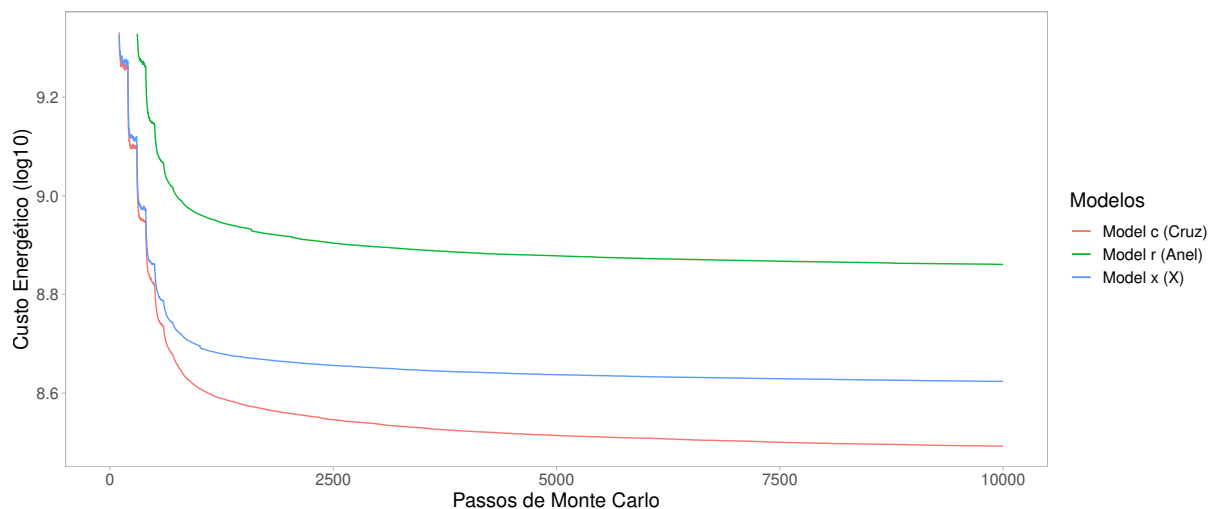


Figura 5.21: Gráfico de Custo Energético para os modelos Anel, Cruz e X na execução de 10000 passos de Monte Carlo.

Com a execução dos passos de Monte Carlo é possível verificarmos a diferença entre os custos energéticos. Até aproximadamente 400 passos os modelos Cruz e X possuem valores próximos, a partir desse valor a diferença torna-se mais explícita. A partir de 7500 passos podemos observar que o decaimento do custo de energia diminui para todos os modelos, inclusive o modelo Cruz que apresentou o menor custo energético para a rede *Homo sapiens*, tornando-se recomendável então a utilização do modelo em Cruz com ao menos 6000 passos para a rede *homo sapiens*. Uma possível explicação para os modelos Cruz e X obterem um menor custo energético poderia ser a quantidade de vizinhos que esses modelos comparam, contudo, ao dividirmos o custo energético do modelo Anel por oito e os modelos Cruz e X por quatro, buscando normalizar a quantidade de comparações, o modelo Anel continua sendo o mais custoso energeticamente. Aumentando a quantidade de passos de Monte Carlo podemos obter um valor menor do custo energético, no entanto, proporcionalmente o tempo de processamento aumentará.

É importante salientarmos que mesmo com a diferença entre custos energéticos entre os modelos é possível notarmos a existência de agrupamentos em pontos semelhantes dentro das modularidades apresentadas. Por exemplo, próximo as posições relativas 0.1, 0.2, 0.35, 0.65 e 0.8, em todos os modelos (Figuras 5.18, 5.19 e 5.20), é possível verificarmos formações *cluster*.

Para a geração da figura 5.21 foram removidos valores negativos do resultado, por exemplo, é possível verificar que a linha de custo energético do modelo Anel tem seu início em valores de passos de Monte Carlo superior aos outros modelos. A geração dos valores negativos deve-se a uma limitação da função responsável por transferir para a linguagem R valores gerados em linguagem C. Com isso, valores de custo maiores que $2,14 \times 10^{+9}$ são impressos no R em valores negativos. Com o andamento dos passos de Monte Carlo o valor energético diminui abaixo desse limite, permitindo assim, mostrar os valores positivos como vemos na figura 5.21.

Com bases nos resultados até aqui, iremos utilizar para os próximos passos o modelo Cruz, que apresentou o menor custo energético, e a janela de tamanho 101, que resultou em agrupamentos de mais fácil identificação.

A figura 5.22 apresenta a posição relativa das proteínas e seus níveis de modularidade. Para demonstrar que os picos desse gráfico formam sub-redes selecionamos trinta e três *clusters*. Na figura 5.23, podemos ver os *clusters* escolhidos com o comprimento de proteínas posicionados dentro dos retângulos coloridos.

As posições relativas, iniciais e finais, de cada proteína para a formação dos *cluster*, além da quantidade de nodos e ligações na formação da sub-rede podem ser observadas na tabela 5.2. Podemos observar que sub-redes selecionadas com base em módulos com pico de valores de modularidade, como por exemplo os módulos D (97 nodos) e AF (139 nodos), apesar de pos-

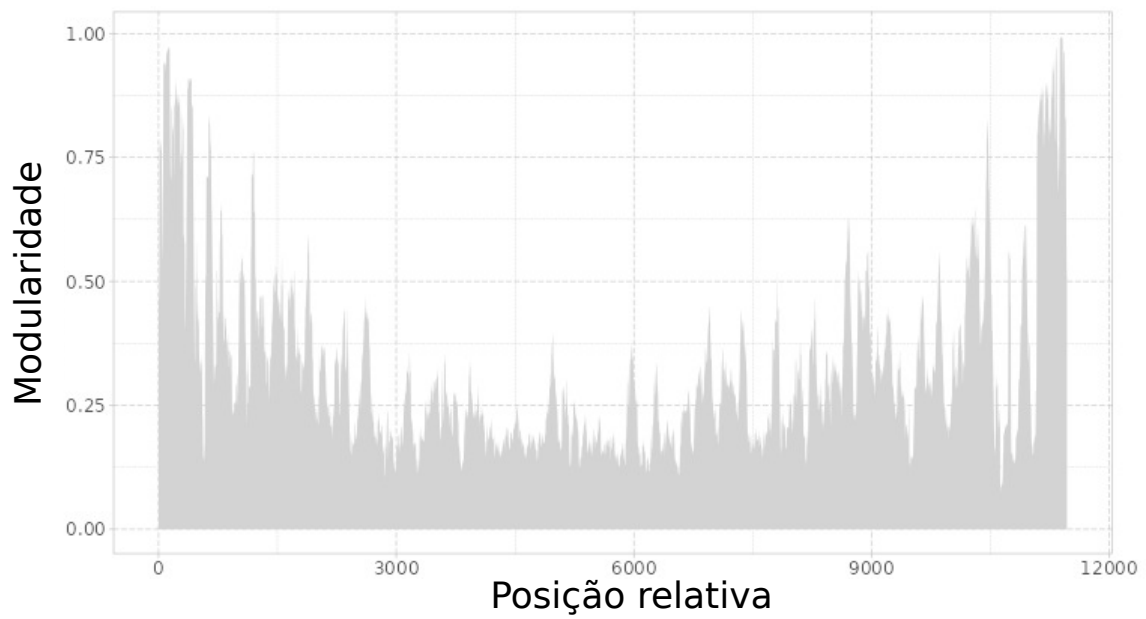


Figura 5.22: Gráfico de modularidade por posição relativa das proteínas. Resultado da aplicação das funções de ordenamento e modularidade nos dados de *Homo Sapiens* obtidos no STRING.

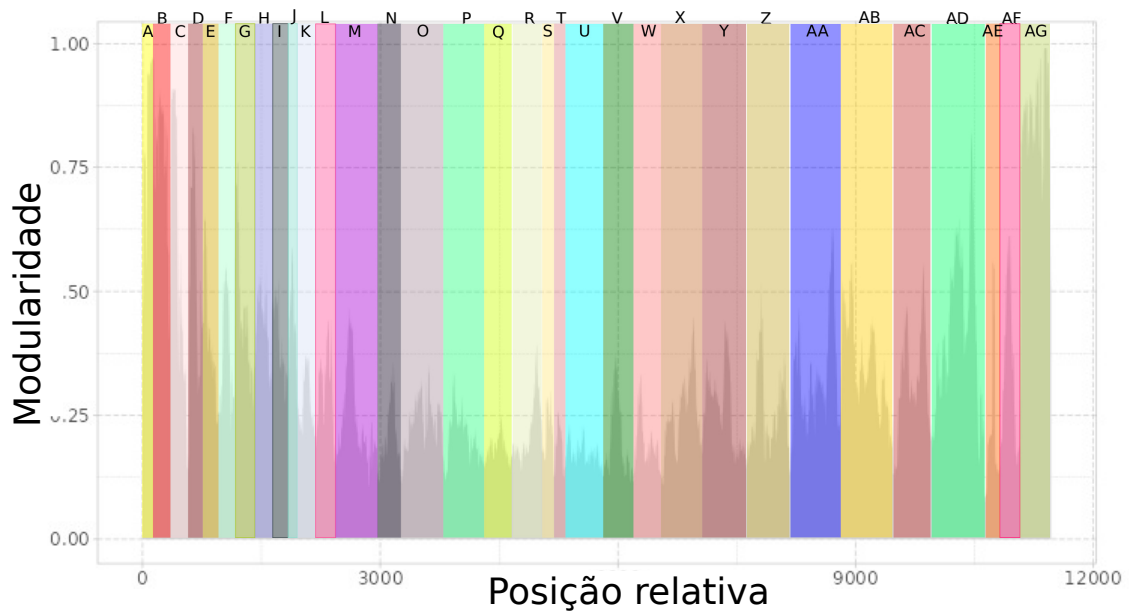


Figura 5.23: Seleção de clusters para criação de sub-redes. As letras representam os trinta e três *clusters* selecionados para exibição de sub-redes.

suírem menos nodos do que o módulo B (142 nodos), possuem grande quantidade de ligações, respectivamente 1.518, 5.776 e 258, demonstrando assim a importância desses sub-grupos para a rede.

Tabela 5.2: Posições relativas das proteínas para formação dos *clusters* selecionados.

Cluster	Posição Inicial	Posição Final	Quantidade de nodos	Quantidade de ligações
A	1	172	164	684
B	173	323	142	258
C	324	551	176	1.594
D	552	703	97	1.518
E	704	950	123	970
F	951	1.139	92	284
G	1.140	1.367	97	694
H	1.368	1.613	136	764
I	1.614	1.784	106	444
J	1.785	2.012	163	786
K	2.013	2.183	127	710
L	2.184	2.391	154	992
M	2.392	2.979	500	11.336
N	2.980	3.264	220	1.722
O	3.265	3.795	492	6.228
P	3.796	4.269	462	7.218
Q	4.270	4.782	513	12.682
R	4.783	5.066	259	3.130
S	5.067	5.218	139	1.226
T	5.219	5.294	70	672
U	5.295	5.806	501	8.792
V	5.807	6.129	309	6.522
W	6.130	6.546	416	7.460
X	6.547	7.039	492	18.860
Y	7.040	7.589	526	11.556
Z	7.590	8.139	434	5.750
AA	8.140	8.784	501	5.626
AB	8.785	9.486	631	9.652
AC	9.487	9.961	436	25.978
AD	9.962	10.606	493	8.498
AE	10.607	10.776	83	270
AF	10.777	11.023	139	5.776
AG	11.024	11.468	419	1.988

Como parte do resultado da aplicação do *workflow* podemos visualizar as sub-redes formadas pelas seleções dos clusters de A a AG no gráfico de modularidade figura 5.24. Com a projeção das sub-redes utilizando o *dashboard* do rTranscriptograma podemos observar o grau de agrupamentos dos clusters selecionados.

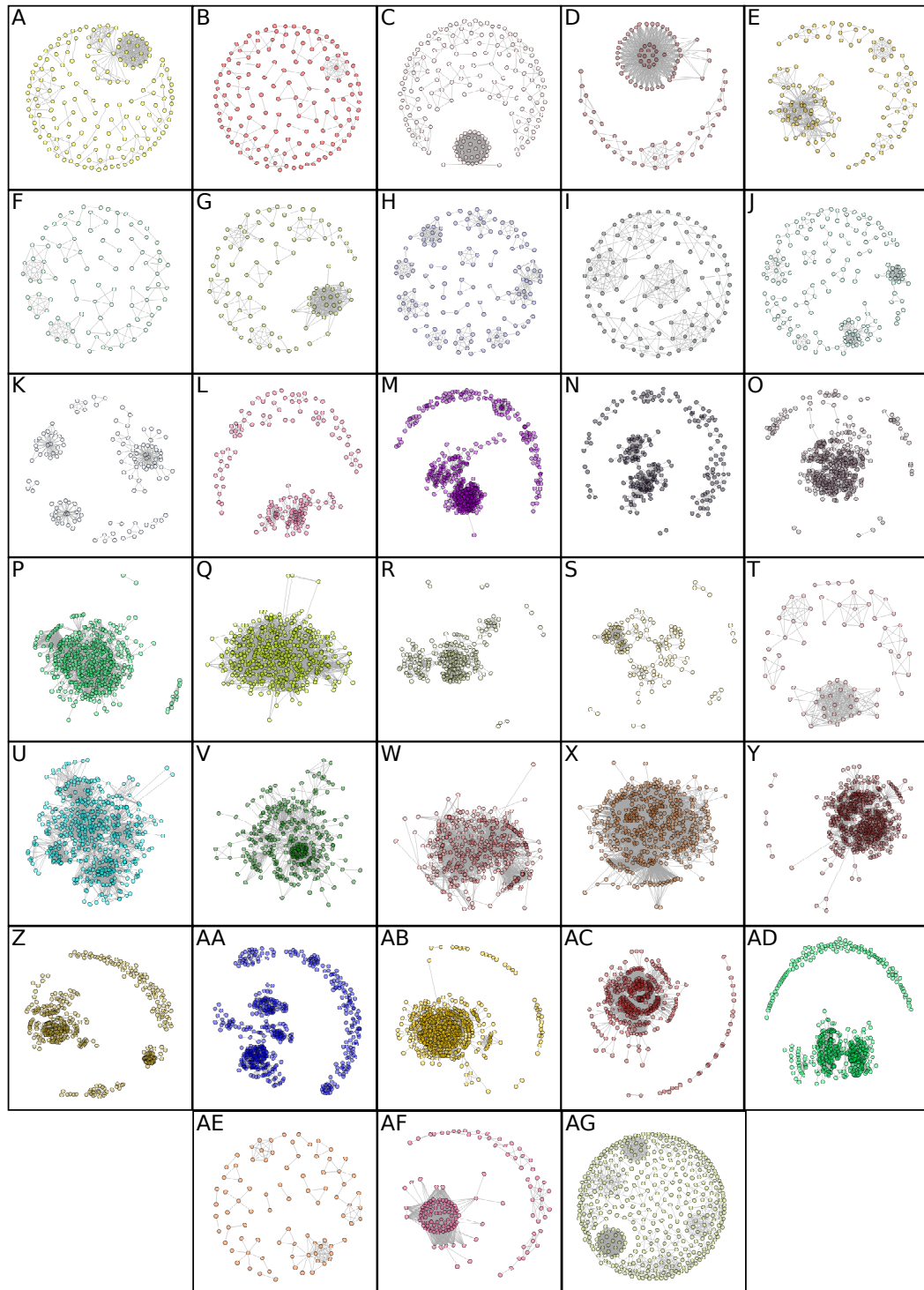


Figura 5.24: Sub-redes criadas a partir da seleção dos módulos. Os gráficos formados representam agrupamentos de interações entre as proteínas da rede do *Homo sapiens*

Para as análises dos dados de expressão foram realizadas três comparações. Abaixo temos a descrição das comparações efetuadas:

- **Segundo conjunto de comparações:**

- **Comparação 1:**

- * **Teste:** - Amostras de fumantes com adenocarcinoma de pulmão
 - * **Controle** - Amostras de pessoas sem adenocarcinoma de pulmão que nunca fumaram

- **Comparação 2:**

- * **Teste** - Amostras de ex-fumantes com adenocarcinoma de pulmão
 - * **Controle** - Amostras de pessoas sem adenocarcinoma de pulmão que nunca fumaram

- **Comparação 3:**

- * **Teste** - Amostras de pessoas com adenocarcinoma de pulmão que nunca fumaram
 - * **Controle** - Amostras de pessoas sem adenocarcinoma de pulmão que nunca fumaram

O estudo de (LANDI et al., 2008) identificou similaridade das assinaturas dos valores de expressão entre as comparações um (amostras de fumantes com adenocarcinoma de pulmão e amostras de pessoas com adenocarcinoma de pulmão que nunca fumaram) e dois (amostras de ex-fumantes com adenocarcinoma de pulmão e amostras de pessoas sem adenocarcinoma de pulmão que nunca fumaram). Segundo o mesmo estudo, alguns indivíduos continuaram apresentando alterações em alguns genes mesmo havendo parado de fumar há mais de 20 anos antes do estudo.

Aplicamos as ferramentas nas series de dados GSE10072 (microarranjo) e GSE81089 (RNA-seq). Na figura 5.25, podemos também observar similaridade e correlação na atividade transcritcional das três comparações aplicadas a série GSE10072.

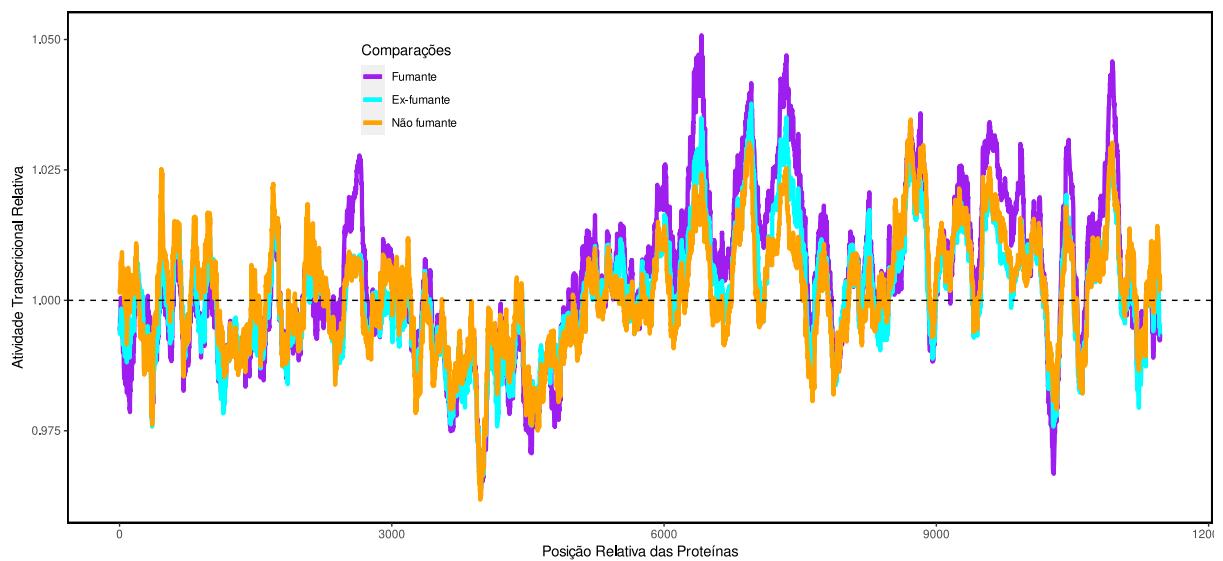


Figura 5.25: Análise da atividade transcricional relativa da série GSE10072 para as comparações de pessoas com adenocarcinoma fumante (cor violeta), ex-fumante (cor ciano) e não fumante (cor mostarda).

No pico próximo às proteínas na posição 2.600 (figura 5.26 (A)) temos, por exemplo, um pico maior para fumantes em relação as outras comparações. Esse pico está associado a algumas ontologias. Por exemplo a GO:0002685 (desenvolvimento de órgãos sensoriais) foi descrito por (SU et al., 2019) como um potencial biomarcador para análise de perfil de pacientes com adenocarcinoma de pulmão. A GO:0009266 (resposta para estímulo de temperatura) está entre dez ontologias de processos biológicos elencados no estudo de (LIU et al., 2020) que se baseia em mineração de dados do TCGA para identificação de assinaturas em câncer de pulmão.

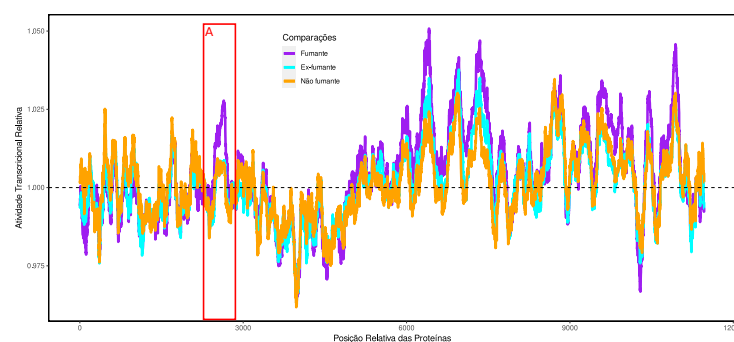


Figura 5.26: Análise da atividade transcricional relativa da série GSE10072 para as comparações de pessoas com adenocarcinoma fumante (cor violeta), ex-fumante (cor ciano) e não fumante (cor mostarda). Na seleção (A) temos um pico maior para fumantes em relação as outras comparações. Esse pico está associado a algumas ontologias como a GO:0002685 (desenvolvimento de órgãos sensoriais) e a GO:0009266 (resposta para estímulo de temperatura).

O diagrama de Venn (figura 5.27) contabiliza a quantidade de ontologias gênicas encontradas em cada comparação para a série GSE10072. A maior concentração de processos alterados

é comum entre as três comparações (653), enquanto que para processos únicos alterados temos para a comparação 1, comparação 2 e comparação 3 respectivamente: 48, 19 e 27.

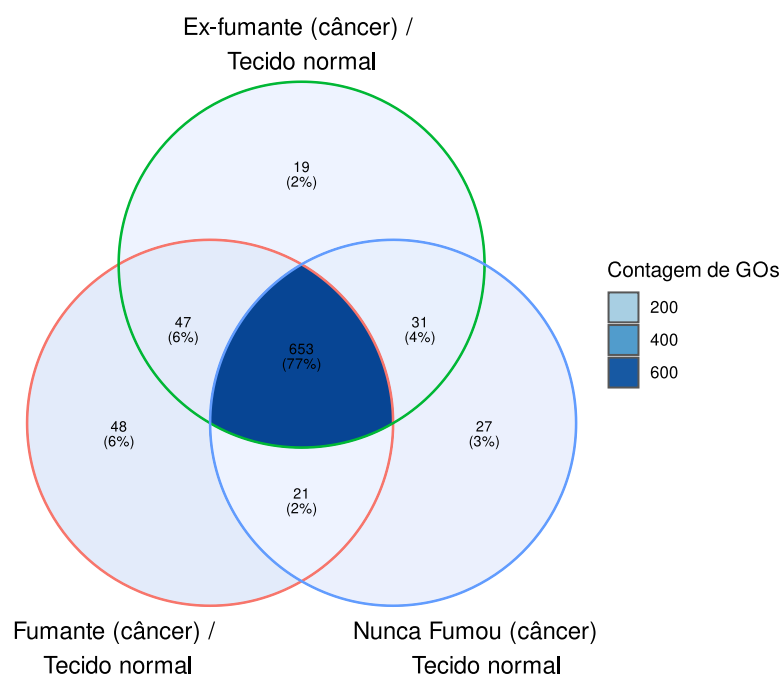


Figura 5.27: Diagrama de Venn para as comparações da série GSE10072 com as contagens de GOs encontradas. A grande maioria de GOs é comum entre os três tipos de câncer (77%). Comuns entre fumantes e ex-fumantes temos 6%. Ontologias comuns entre ex-fumantes e amostras de pessoas que nunca fumaram representam 4%. GOs comuns entre amostras de pessoas que nunca fumaram e fumantes representam 2%. Ontologias que não aparecem em outros grupos temos para fumantes, ex-fumantes e que nunca fumaram, respectivamente: 6%, 2% e 3%.

Para exemplificar a utilização da ferramenta *levi* selecionamos o *cluster* AC (tabela 5.2) que possui 436 nodos e 25.978 ligações e aplicamos sobre a rede os dados de expressão de pessoas fumantes com adenocarcinoma da série GSE10072. Entre as proteínas superexpressas nesse cluster temos ENSP00000302896, ENSP00000472985, ENSP00000341885, ENSP00000230050, ENSP00000331019, ENSP00000346050, ENSP00000379339, ENSP00000245539, ENSP00000393241 e ENSP00000429374 que estão localizadas na área com maior temperatura (vermelho) na sobreposição dos dados de expressão sobre a rede (figura 5.28).

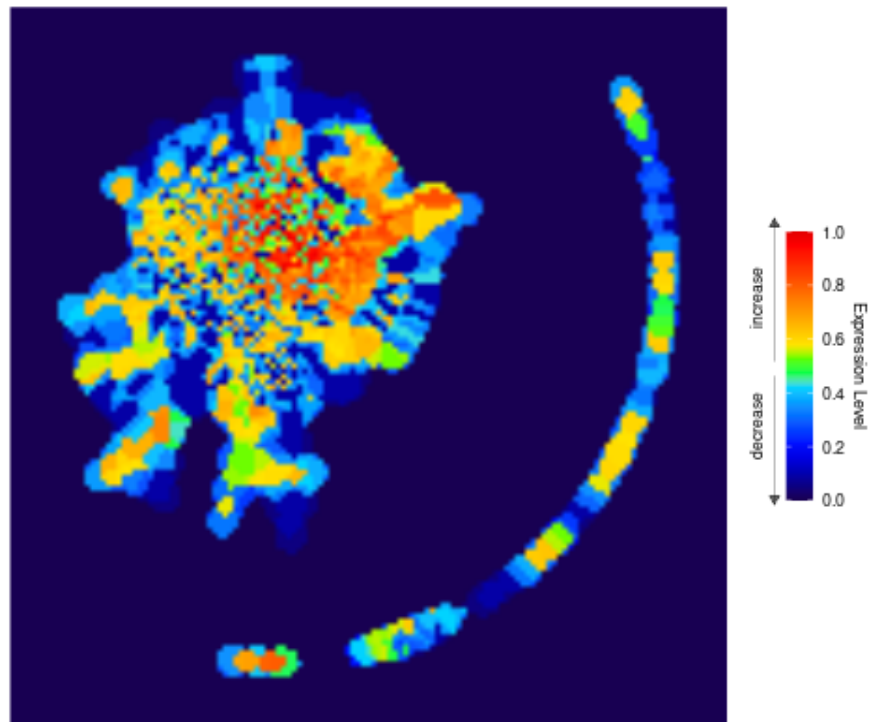


Figura 5.28: Aplicação do *levi* para visualização da sobreposição do cluster AC em dados de expressão da série GSE10072. O mapa de calor (*heatmap*) visualizado como resultado do *levi* demonstra os níveis de expressão gênica projetado na rede de interação proteica. Áreas de expressão vermelhas são superexpressas, enquanto, áreas de expressão azul são subexpressas. Com os dados de expressão gênica da séries GSE10072 podemos observar um nível de expressão maior na área central do agrupamento AC onde temos as proteínas ENSP00000302896, ENSP00000472985, ENSP00000341885, ENSP00000230050, ENSP00000331019, ENSP00000346050, ENSP00000379339, ENSP00000245539, ENSP00000393241 e ENSP00000429374.

Para a segunda série de dados de expressão GSE81089 foram criados também os gráficos de atividade transcricional relativa por posição relativa das proteínas (5.29) e o diagrama de Venn (5.31).

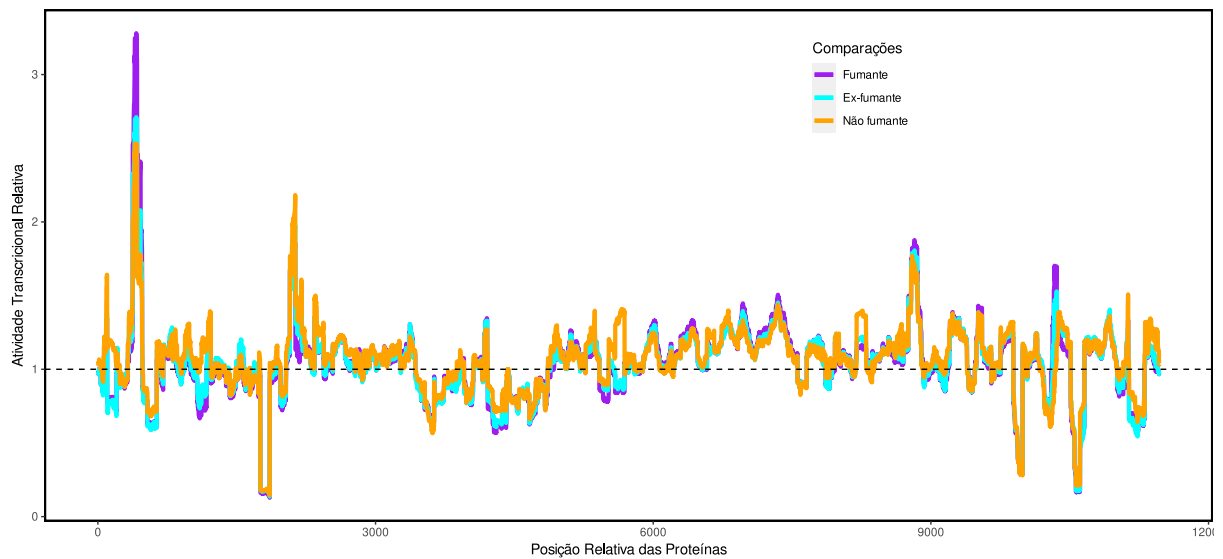


Figura 5.29: Análise da atividade transcricional relativa da série GSE10072 para as comparações de pessoas com adenocarcinoma fumante (cor violeta), ex-fumante (cor ciano) e não fumante (cor mostarda).

Assim como observado com a série de dados GSE10072, existe um alto grau de similaridades entre as três comparações. Próximo ao valor 500 (figura 5.30 (A)) da posição relativa das proteínas é possível observar um pico maior de fumantes em relação as outras três comparações. Esse pico está associado a ontologias como a GO:0010035 (Resposta à substância inorgânica) e GO:0016032 (processo viral), sendo que a primeira dessas ontologias também foi encontradas na análise do enriquecimento do estudo de (LIU et al., 2020) que buscou encontrar biomarcadores para adenocarcinoma de pulmão e a segunda ontologia está entre os dez processos mais enriquecidos no estudo de (LIU et al., 2016) que fez uma análise sistemática em perfis de expressão relacionados a metastase de câncer de pulmão.

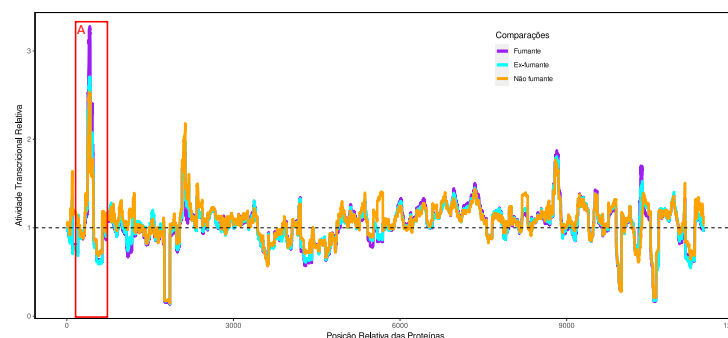


Figura 5.30: Análise da atividade transcricional relativa da série GSE10072 para as comparações de pessoas com adenocarcinoma fumante (cor violeta), ex-fumante (cor ciano) e não fumante (cor mostarda). Na seleção (A) temos um pico maior para fumantes em relação as outras comparações. Esse pico está associado a algumas ontologias como a GO:0010035 (Resposta à substância inorgânica) e GO:0016032 (processo viral).

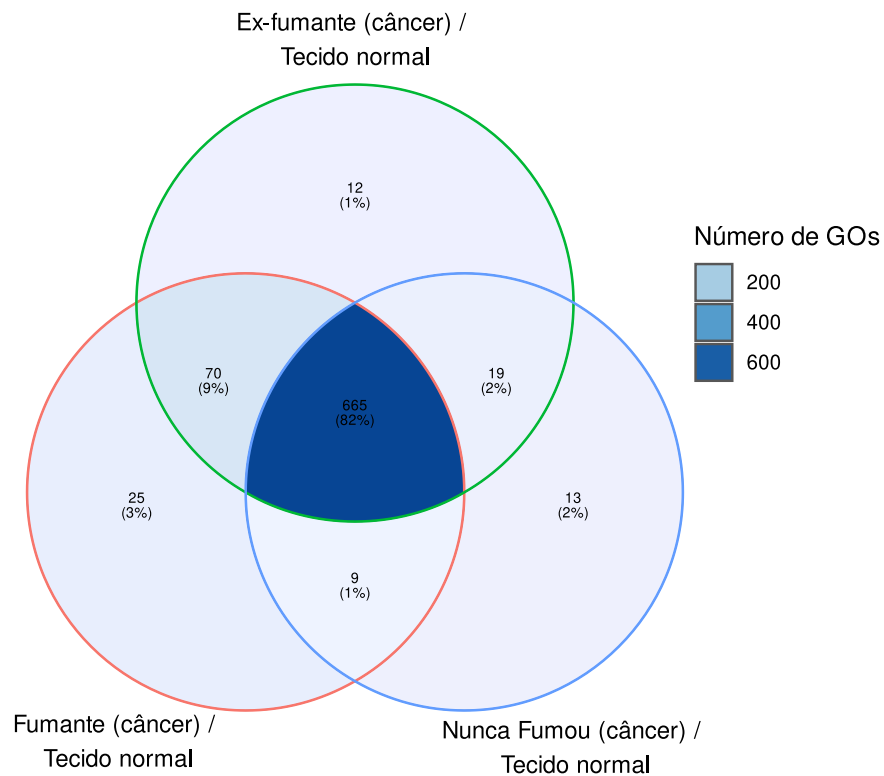


Figura 5.31: Diagrama de Venn para as comparações da série GSE81089 com as contagens de GOs encontradas. A grande maioria de GOs é comum entre os três tipos de câncer (82%). Comuns entre fumantes e ex-fumantes temos 9%. Ontologias comuns entre ex-fumantes e amostras de pessoas que nunca fumaram representam 2%. GOs comuns entre amostras de pessoas que nunca fumaram e fumantes representam 1%. Ontologias que não aparecem em outros grupos temos para fumantes, ex-fumantes e que nunca fumaram, respectivamente: 3%, 1% e 2%.

Utilizamos o mesmo *cluster* AC (tabela 5.2) para a sobreposição dos dados de expressão da série GSE81089 com os dados de expressão de pessoas fumantes com adenocarcinoma (figura 5.32), podemos verificar uma maior centralização das proteínas superexpressas no centro do agrupamento e mais regiões subexpressas ao redor do *cluster*.

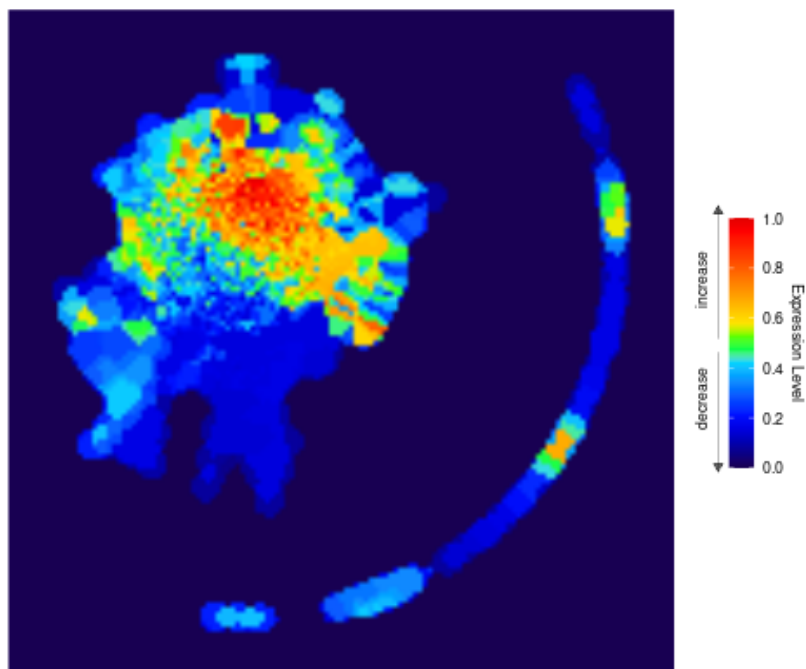


Figura 5.32: Aplicação do *levi* para visualização da sobreposição do cluster AC em dados de expressão da série GSE81089. O mapa de calor (*heatmap*) visualizado como resultado do *levi* demonstra os níveis de expressão gênica projetado na rede de interação proteica. Áreas de expressão vermelhas são superexpressas, enquanto, áreas de expressão azul são subexpressas. Com os dados de expressão gênica da séries GSE81089 podemos observar um nível de expressão maior na área central do agrupamento AC onde temos as proteínas ENSP00000302896, ENSP00000472985, ENSP00000341885, ENSP00000230050, ENSP00000331019, ENSP00000346050, ENSP00000379339, ENSP00000245539, ENSP00000393241 e ENSP00000429374.

As tabelas 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, apresentaram amostras das GOs e seus p-valores, p-ajustados, q-valores, módulos e razão, para as comparações 1 (indivíduos com câncer fumantes), comparação 2 (indivíduos com câncer ex-fumantes) e comparação 3 (indivíduos com câncer que nunca fumaram) da série GSE10072 e as comparações comparações 1 (indivíduos com câncer fumantes), comparação 2 (indivíduos com câncer ex-fumantes) e comparação 3 (indivíduos com câncer que nunca fumaram) da série GSE81089 respectivamente. Esses dados são utilizados para a construção do transcriptograma.

Os arquivos com os dados das GOs exibidas nos diagramas de Venn, assim como os arquivos contendo os dados detalhados dos p-valores, p-ajustados, q-valores, módulos e razão podem ser acessados pelo link (**material suplementar**).

Tabela 5.3: Amostra de processos biológicos existentes no *cluster* 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer fumantes da série GSE10072.

GO	Descrição	P-valor	P-ajustado	Q-valor	Módulo	Razão
GO:0030216	keratinocyte differentiation	3,55E-49	7,75E-47	7,17E-47	cluster_1	0.6
GO:0070193	synaptonemal complex organization	1,06E-02	3,87E-01	3,58E-01	cluster_1	0.5
GO:0045143	homologous chromosome segregation	2,44E-02	4,10E-01	3,79E-01	cluster_1	0.5
GO:0032527	protein exit from endoplasmic reticulum	3,14E-02	4,10E-01	3,79E-01	cluster_1	1
GO:0015711	organic anion transport	4,84E-02	4,10E-01	3,79E-01	cluster_1	0.6
GO:0010560	positive regulation of glycoprotein biosynthetic process	4,97E-02	4,10E-01	3,79E-01	cluster_1	0.5
GO:0010866	regulation of triglyceride biosynthetic process	5,27E-02	4,19E-01	3,87E-01	cluster_1	0.5
GO:0055081	anion homeostasis	5,27E-02	4,19E-01	3,87E-01	cluster_1	0.5
GO:0032782	bile acid secretion	5,57E-02	4,20E-01	3,88E-01	cluster_1	0.5
GO:2000677	regulation of transcription regulatory region DNA binding	7,07E-02	4,44E-01	4,11E-01	cluster_1	0.5
GO:1903019	negative regulation of glycoprotein metabolic process	7,66E-02	4,44E-01	4,11E-01	cluster_1	0.5
GO:0048608	reproductive structure development	7,81E-02	4,44E-01	4,11E-01	cluster_1	1
GO:0072079	nephron tubule formation	8,54E-02	4,44E-01	4,11E-01	cluster_1	1
GO:0019432	triglyceride biosynthetic process	9,41E-02	4,62E-01	4,27E-01	cluster_1	0.5
GO:0072210	metanephric nephron development	9,70E-02	4,66E-01	4,31E-01	cluster_1	1
GO:0060415	muscle tissue morphogenesis	1,17E-01	4,80E-01	4,44E-01	cluster_1	0.5
GO:0008585	female gonad development	1,28E-01	4,93E-01	4,56E-01	cluster_1	0.5
GO:0072028	nephron morphogenesis	1,87E-01	5,35E-01	4,95E-01	cluster_1	1
GO:1905477	positive regulation of protein localization to membrane	2,10E-01	5,54E-01	5,12E-01	cluster_1	1
GO:0045639	positive regulation of myeloid cell differentiation	2,13E-01	5,57E-01	5,15E-01	cluster_1	1

Tabela 5.4: Amostra de processos biológicos existentes no *cluster* 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer ex-fumantes da série GSE10072.

GO	Descrição	P-valor	P-ajustado	Q-valor	Módulo	Razão
GO:0015711	organic anion transport	2,64E-03	7,45E-02	5,93E-02	cluster_1	0,6
GO:0032527	protein exit from endoplasmic reticulum	1,09E-02	1,47E-01	1,17E-01	cluster_1	1
GO:0060563	neuroepithelial cell differentiation	1,52E-02	1,47E-01	1,17E-01	cluster_1	0,5
GO:0010560	positive regulation of glycoprotein biosynthetic process	1,74E-02	1,47E-01	1,17E-01	cluster_1	0,5
GO:0055081	anion homeostasis	1,85E-02	1,51E-01	1,20E-01	cluster_1	0,5
GO:0032782	bile acid secretion	1,96E-02	1,51E-01	1,20E-01	cluster_1	0,5
GO:0070169	positive regulation of biomineral tissue development	3,03E-02	1,64E-01	1,30E-01	cluster_1	1
GO:0072079	nephron tubule formation	3,03E-02	1,64E-01	1,30E-01	cluster_1	1
GO:0072210	metanephric nephron development	3,45E-02	1,65E-01	1,32E-01	cluster_1	1
GO:0072028	nephron morphogenesis	6,89E-02	2,04E-01	1,63E-01	cluster_1	1
GO:1905477	positive regulation of protein localization to membrane	7,81E-02	2,13E-01	1,70E-01	cluster_1	1
GO:0061515	myeloid cell development	9,62E-02	2,17E-01	1,73E-01	cluster_1	0,5
GO:0006575	cellular modified amino acid metabolic process	1,80E-01	2,96E-01	2,36E-01	cluster_1	1
GO:0050953	sensory perception of light stimulus	1,92E-01	3,01E-01	2,39E-01	cluster_1	0,5
GO:0031098	stress-activated protein kinase signaling cascade	2,13E-01	3,14E-01	2,50E-01	cluster_1	1
GO:0042063	gliogenesis	2,61E-01	3,47E-01	2,76E-01	cluster_1	1
GO:0060562	epithelial tube morphogenesis	2,67E-01	3,51E-01	2,80E-01	cluster_1	1
GO:0045862	positive regulation of proteolysis	2,90E-01	3,63E-01	2,89E-01	cluster_1	1
GO:0048732	gland development	3,05E-01	3,74E-01	2,97E-01	cluster_1	0,5
GO:0048880	sensory system development	3,11E-01	3,74E-01	2,98E-01	cluster_1	1

Tabela 5.5: Amostra de processos biológicos existentes no *cluster* 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer que nunca fumaram da série GSE10072.

GO	Descrição	P-valor	P-ajustado	Q-valor	Módulo	Razão
GO:0010866	regulation of triglyceride biosynthetic process	1,02E-02	1,42E-01	1,06E-01	cluster_1	0,5
GO:0019432	triglyceride biosynthetic process	1,85E-02	1,42E-01	1,06E-01	cluster_1	0,5
GO:0046890	regulation of lipid biosynthetic process	5,24E-02	1,58E-01	1,19E-01	cluster_1	1
GO:0071453	cellular response to oxygen levels	7,00E-02	1,58E-01	1,19E-01	cluster_1	0,5
GO:0048762	mesenchymal cell differentiation	8,35E-02	1,67E-01	1,26E-01	cluster_1	1
GO:0006575	cellular modified amino acid metabolic process	1,03E-01	1,82E-01	1,36E-01	cluster_1	1
GO:0060485	mesenchyme development	1,07E-01	1,82E-01	1,36E-01	cluster_1	0,5
GO:0050792	regulation of viral process	1,14E-01	1,84E-01	1,38E-01	cluster_1	0,5
GO:0031098	stress-activated protein kinase signaling cascade	1,23E-01	1,86E-01	1,40E-01	cluster_1	1
GO:0051701	biological process involved in interaction with host	1,48E-01	1,93E-01	1,45E-01	cluster_1	1
GO:0045862	positive regulation of proteolysis	1,71E-01	2,08E-01	1,56E-01	cluster_1	1
GO:0016032	viral process	2,46E-01	2,57E-01	1,93E-01	cluster_1	1

Tabela 5.6: Amostra de processos biológicos existentes no *cluster* 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer fumantes da série GSE81089.

GO	Descrição	P-valor	P-ajustado	Q-valor	Módulo	Razão
GO:0032527	protein exit from endoplasmic reticulum	1,25E-02	1,18E-01	8,86E-02	cluster_1	1
GO:0010560	positive regulation of glycoprotein biosynthetic process	2,00E-02	1,18E-01	8,86E-02	cluster_1	0.5
GO:0010866	regulation of triglyceride biosynthetic process	2,12E-02	1,21E-01	9,03E-02	cluster_1	0.5
GO:0032782	bile acid secretion	2,25E-02	1,23E-01	9,19E-02	cluster_1	0.5
GO:0019432	triglyceride biosynthetic process	3,84E-02	1,51E-01	1,13E-01	cluster_1	0.5
GO:0048608	reproductive structure development	5,89E-02	1,85E-01	1,38E-01	cluster_1	0.6
GO:1905477	positive regulation of protein localization to membrane	8,93E-02	2,13E-01	1,60E-01	cluster_1	1
GO:0046890	regulation of lipid biosynthetic process	1,06E-01	2,31E-01	1,73E-01	cluster_1	1
GO:0071453	cellular response to oxygen levels	1,41E-01	2,56E-01	1,92E-01	cluster_1	0.5
GO:0007127	meiosis I	1,42E-01	2,56E-01	1,92E-01	cluster_1	1
GO:0048762	mesenchymal cell differentiation	1,67E-01	2,75E-01	2,06E-01	cluster_1	1
GO:0060485	mesenchyme development	2,10E-01	3,11E-01	2,33E-01	cluster_1	0.5
GO:0050953	sensory perception of light stimulus	2,17E-01	3,11E-01	2,33E-01	cluster_1	0.5
GO:0048880	sensory system development	3,48E-01	3,86E-01	2,89E-01	cluster_1	1

Tabela 5.7: Amostra de processos biológicos existentes no *cluster* 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer ex-fumantes da série GSE81089.

GO	Descrição	P-valor	P-ajustado	Q-valor	Módulo	Razão
GO:0032527	protein exit from endoplasmic reticulum	1,36E-02	1,25E-01	9,35E-02	cluster_1	1
GO:0010560	positive regulation of glycoprotein biosynthetic process	2,17E-02	1,25E-01	9,35E-02	cluster_1	0.5
GO:0010866	regulation of triglyceride biosynthetic process	2,31E-02	1,25E-01	9,35E-02	cluster_1	0.5
GO:0032782	bile acid secretion	2,44E-02	1,27E-01	9,54E-02	cluster_1	0.5
GO:0019432	triglyceride biosynthetic process	4,17E-02	1,60E-01	1,20E-01	cluster_1	0.5
GO:0048608	reproductive structure development	6,83E-02	2,04E-01	1,53E-01	cluster_1	0.6
GO:1905477	positive regulation of protein localization to membrane	9,67E-02	2,26E-01	1,70E-01	cluster_1	1
GO:0046890	regulation of lipid biosynthetic process	1,15E-01	2,46E-01	1,85E-01	cluster_1	1
GO:0071453	cellular response to oxygen levels	1,52E-01	2,74E-01	2,05E-01	cluster_1	0.5
GO:0007127	meiosis I	1,53E-01	2,74E-01	2,05E-01	cluster_1	1
GO:0048762	mesenchymal cell differentiation	1,80E-01	2,95E-01	2,21E-01	cluster_1	1
GO:0060485	mesenchyme development	2,26E-01	3,30E-01	2,48E-01	cluster_1	0.5
GO:0050953	sensory perception of light stimulus	2,34E-01	3,31E-01	2,48E-01	cluster_1	0.5
GO:0048880	sensory system development	3,72E-01	4,11E-01	3,08E-01	cluster_1	1

Tabela 5.8: Amostra de processos biológicos existentes no *cluster* 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer que nunca fumaram da série GSE81089.

GO	Descrição	P-valor	P-ajustado	Q-valor	Módulo	Razão
GO:0015711	organic anion transport	1,11E-03	7,01E-02	4,74E-02	cluster_1	0.6
GO:0032527	protein exit from endoplasmic reticulum	8,19E-03	7,01E-02	4,74E-02	cluster_1	1
GO:0009913	epidermal cell differentiation	9,24E-03	7,01E-02	4,74E-02	cluster_1	1
GO:0010560	positive regulation of glycoprotein biosynthetic process	1,31E-02	7,01E-02	4,74E-02	cluster_1	0.5
GO:0055081	anion homeostasis	1,39E-02	7,25E-02	4,90E-02	cluster_1	0.5
GO:0032782	bile acid secretion	1,47E-02	7,28E-02	4,92E-02	cluster_1	0.5
GO:0072079	nephron tubule formation	2,28E-02	9,16E-02	6,20E-02	cluster_1	1
GO:0072210	metanephric nephron development	2,60E-02	9,56E-02	6,47E-02	cluster_1	1
GO:0048608	reproductive structure development	2,68E-02	9,56E-02	6,47E-02	cluster_1	0.6
GO:0072028	nephron morphogenesis	5,22E-02	1,31E-01	8,86E-02	cluster_1	1
GO:1905477	positive regulation of protein localization to membrane	5,92E-02	1,39E-01	9,42E-02	cluster_1	1
GO:0006575	cellular modified amino acid metabolic process	1,38E-01	2,05E-01	1,39E-01	cluster_1	1
GO:0031098	stress-activated protein kinase signaling cascade	1,64E-01	2,23E-01	1,51E-01	cluster_1	1
GO:0042063	gliogenesis	2,03E-01	2,48E-01	1,68E-01	cluster_1	1
GO:0060562	epithelial tube morphogenesis	2,08E-01	2,51E-01	1,70E-01	cluster_1	1
GO:0045862	positive regulation of proteolysis	2,26E-01	2,65E-01	1,79E-01	cluster_1	1
GO:0048732	gland development	2,39E-01	2,75E-01	1,86E-01	cluster_1	0.5

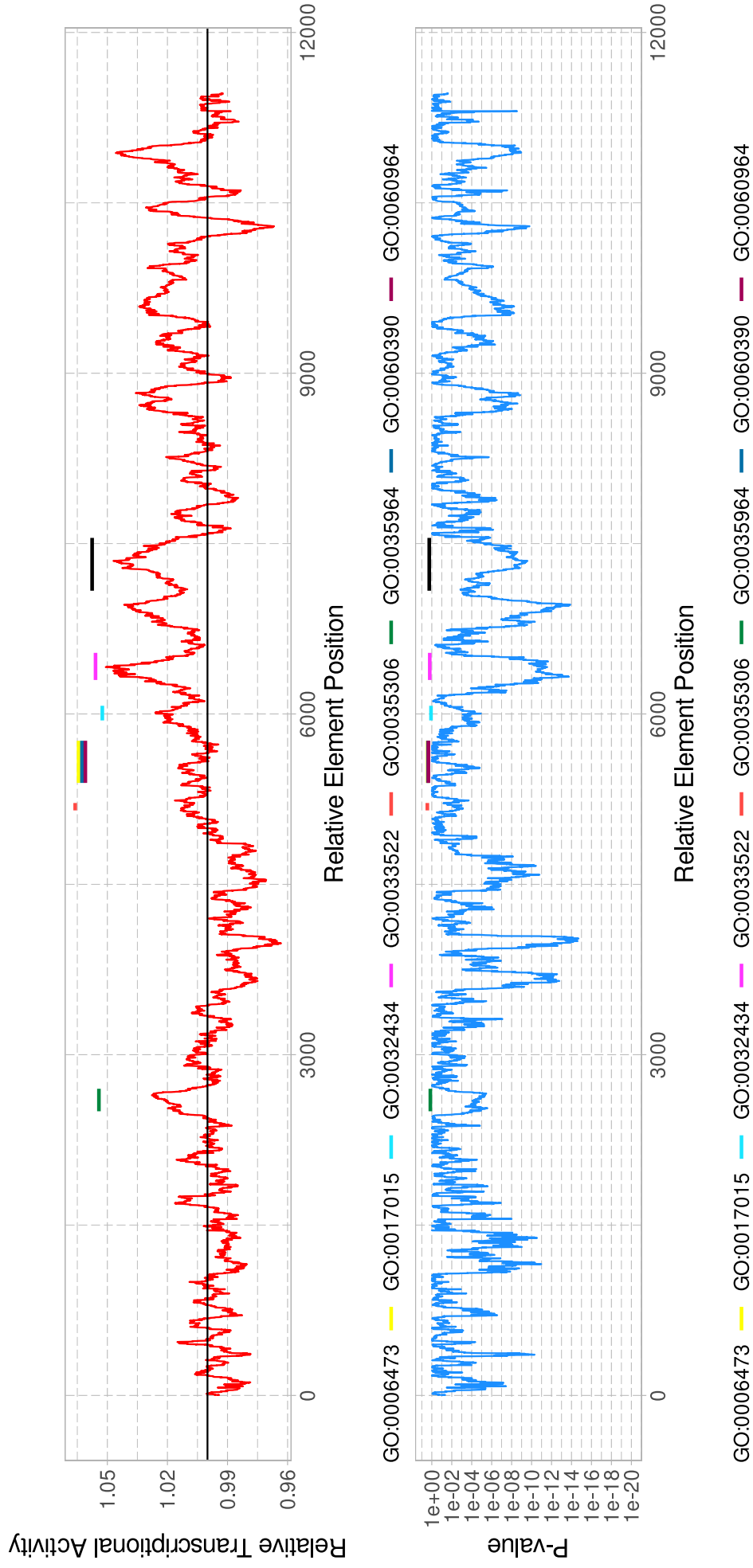


Figura 5.33: Transcriptograma da comparação 1 (fumantes com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória.

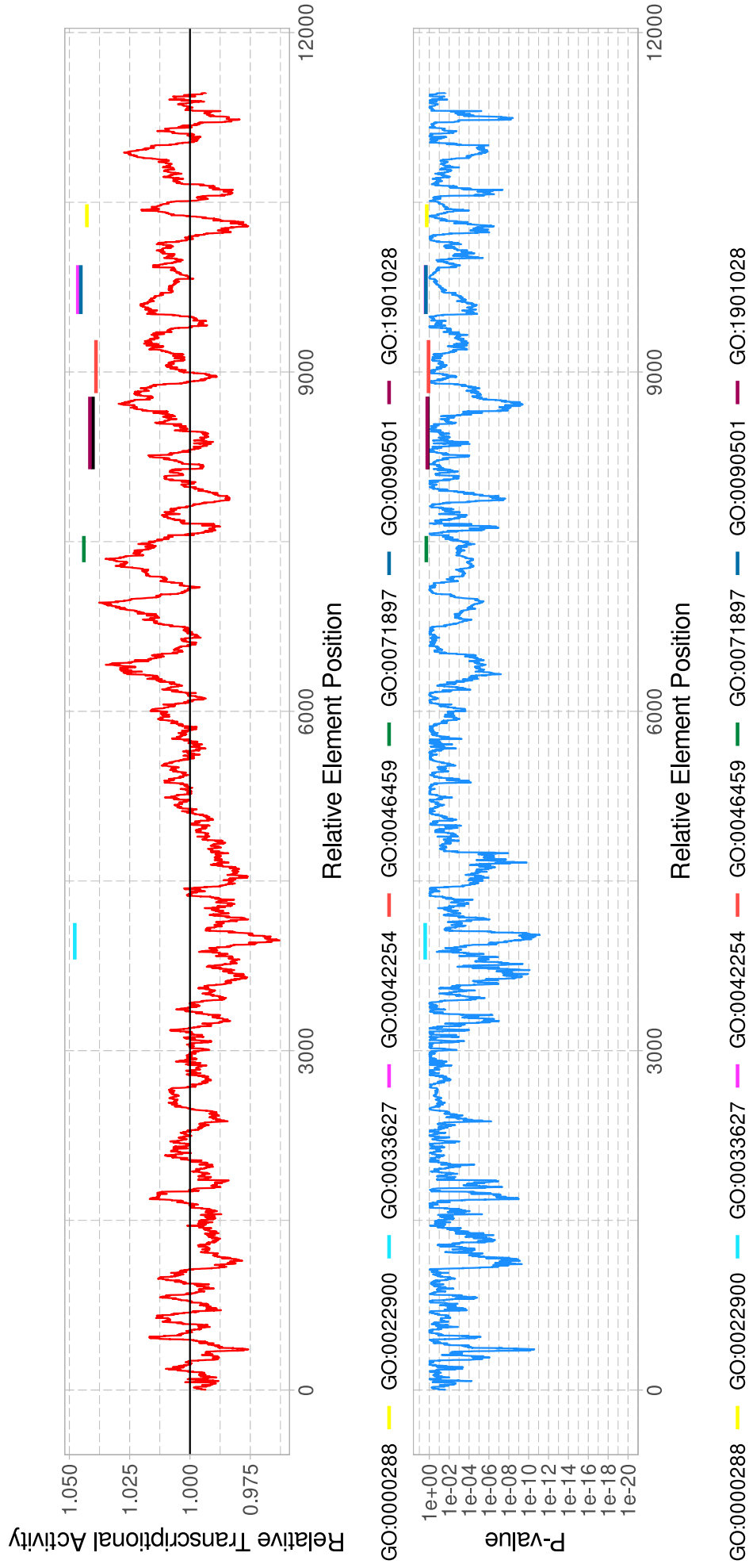


Figura 5.34: Transcriptograma da comparação da comparação 2 (ex-fumantes com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória.

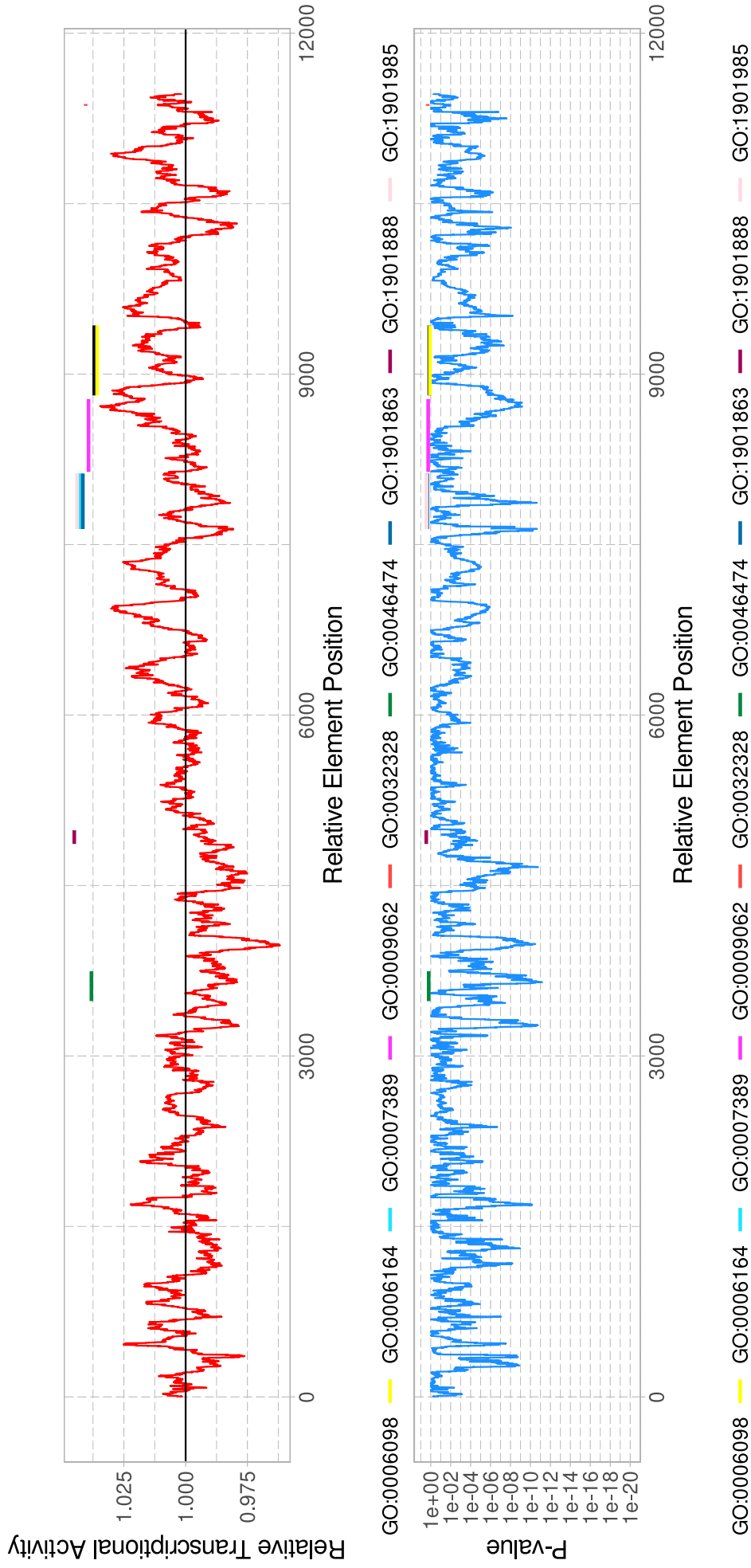


Figura 5.35: Transcriptograma da comparação 3 (indivíduos que nunca fumaram com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória.

Com os dados de expressão processados, os valores de *fold-change* e p-valores é possível a construção dos gráficos do transcriptograma.

Em nossos resultados podemos observar que todos os genes existentes na rede de proteínas foram utilizados na análise, diferente dos resultados de (LANDI et al., 2008) que removeram genes que não se enquadravam nos filtros de *fold-change* mínimo de 1,5 para os genes superexpressos e o valor máximo de 0.6667 para subexpressos. Em seu estudo, (LANDI et al., 2008), utilizou três possíveis comparações, entre amostras com adenocarcinoma de indivíduos fumantes e que nunca fumaram (comparação 1), entre amostras com adenocarcinoma de indivíduos ex-fumantes e que nunca fumaram (comparação 2) e entre amostras de indivíduos com adenocarcinoma e amostras de indivíduos com tecido normal (comparação 3). Em sua análise foi possível observar os níveis de expressão principalmente entre fumantes e ex-fumantes. Nas nossas comparações, utilizamos como referência amostras de tecidos normais e, com isso, pudemos observar os dados de expressão para pessoas com câncer que nunca fumaram. Outro fator importante é que em nossos resultados obtivemos um p-valor mínimo de 1×10^{-14} enquanto que (LANDI et al., 2008) obteve um p-valor de no mínimo de 1×10^{-3} .

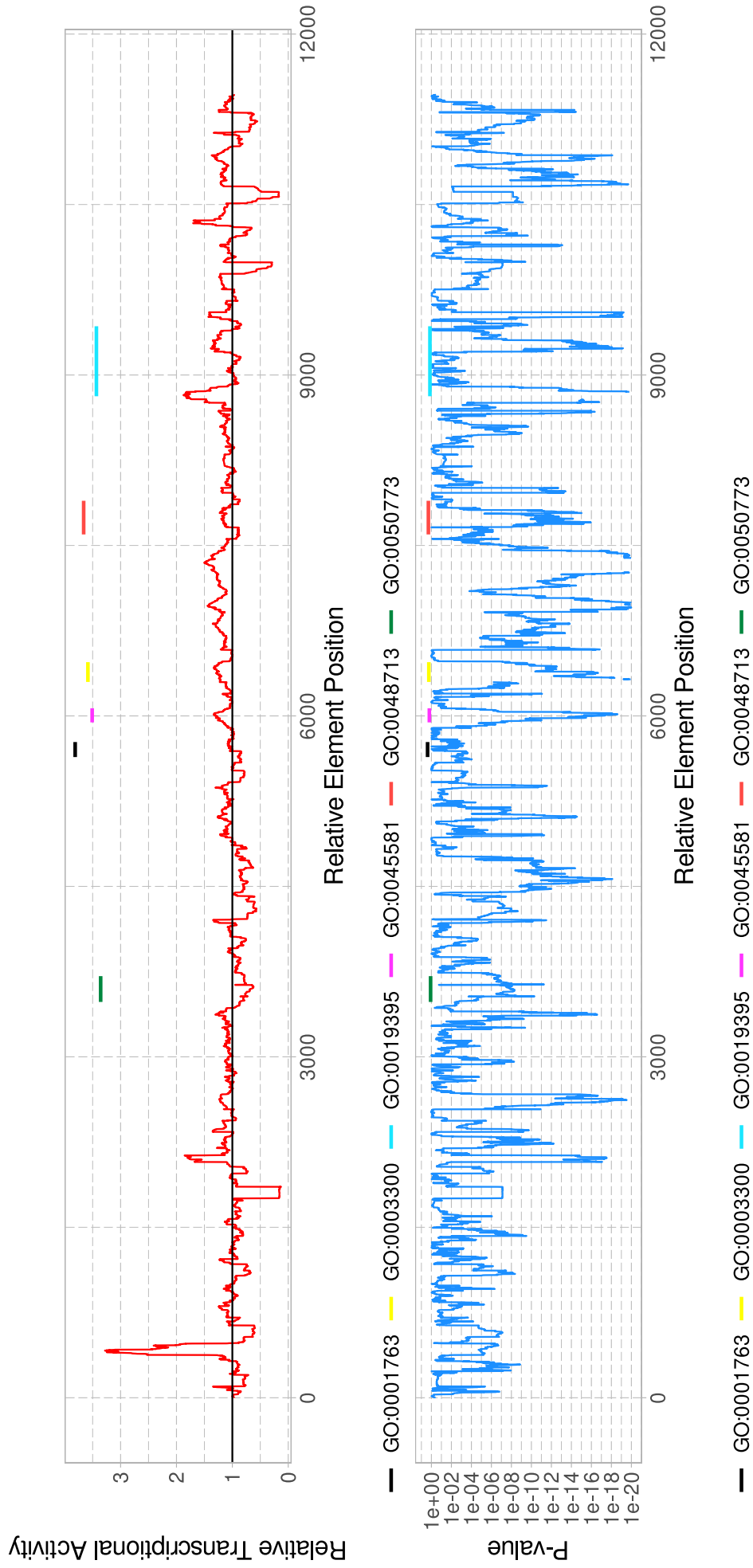


Figura 5.36: Transcriptograma da comparação 1 (fumantes com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória.

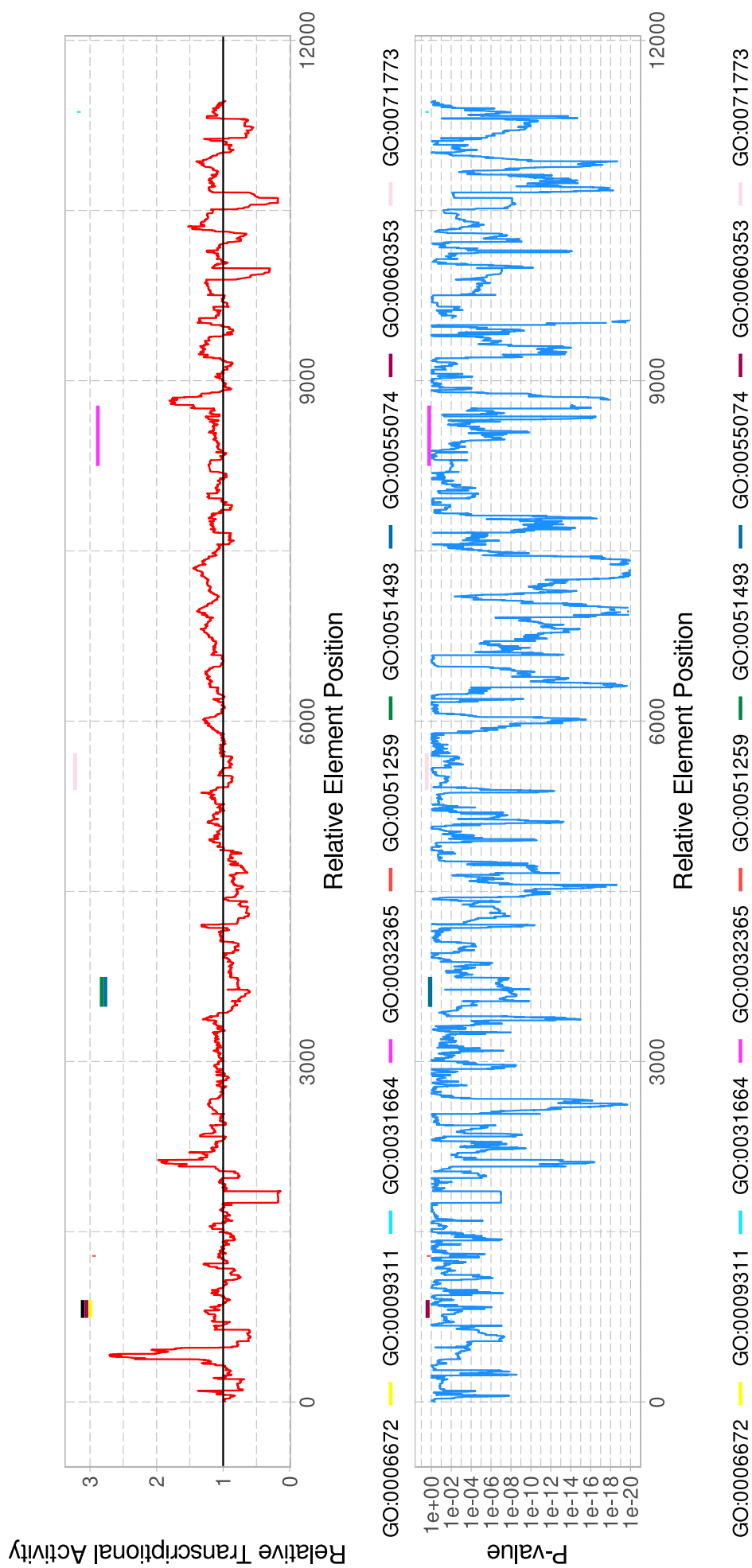


Figura 5.37: Transcriptograma da comparação da comparação 2 (ex-fumantes com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória.

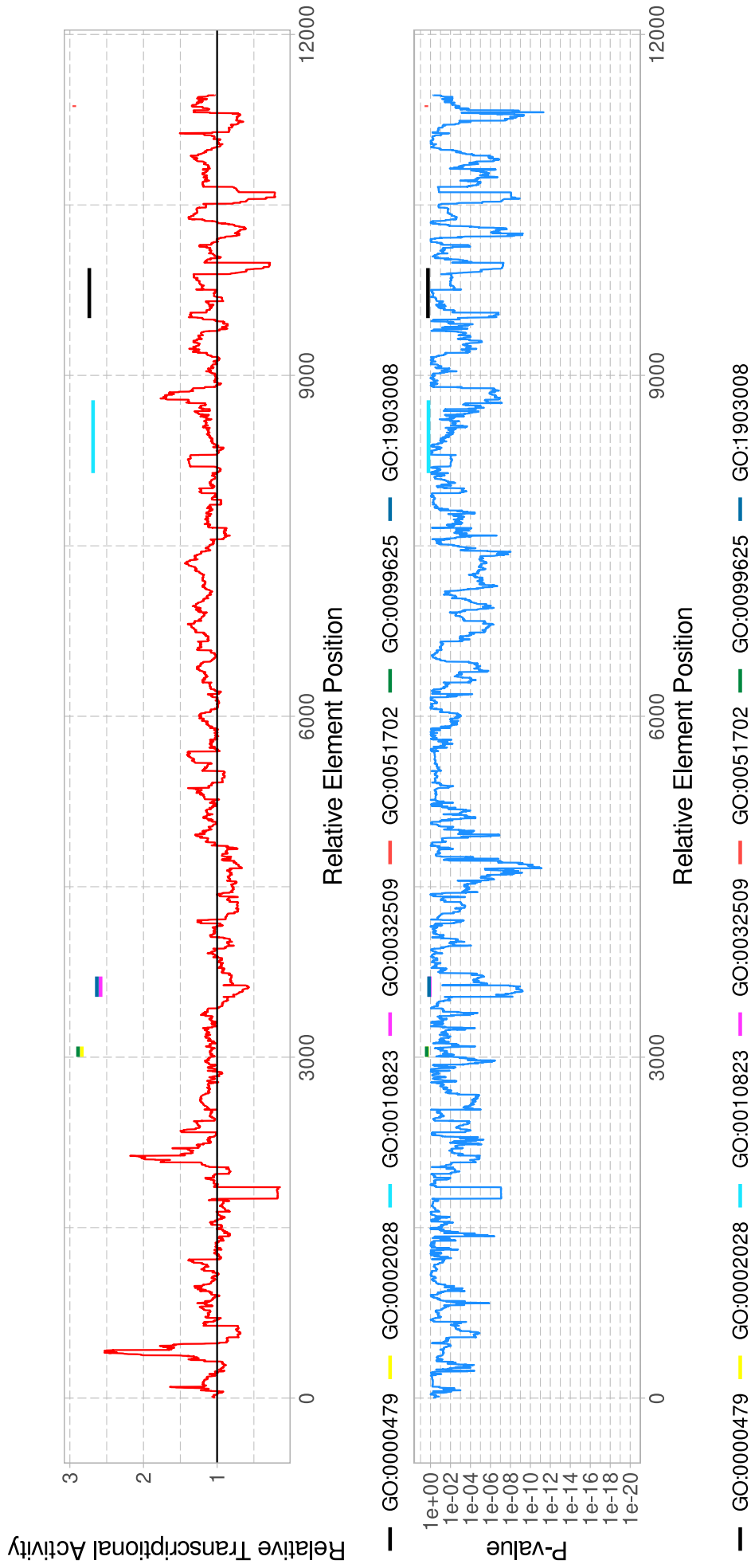


Figura 5.38: Transcriptograma da comparação 3 (indivíduos que nunca fumaram com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória.

Em seus resultados (GOLDMANN et al., 2021) descarta genes que não se enquadram nos filtros de p-valor ajustado menor que 0.05, limitando a quantidade de genes identificados. Na aplicação do rTranscriptograma utilizamos como base toda a rede de proteínas permitindo assim a visualização de maior quantidade de processos biológicos. Outro resultado interessante em nossa análise é que os p-valores para esse conjunto de dados de expressão obtiveram valores mínimos de p-valor de 1×10^{-18} . Em seu trabalho, (GOLDMANN et al., 2021), utiliza a série GSE81089 para verificar os dados de expressão de pessoas com adenocarcinoma em comparação a série GSE37745 com dados de expressão de amostras com carcinoma de células escamosas, outra subclassificação das câncer de pulmão de células não pequenas *Nons-Small Cell Lung Cancer*. Como em nossos estudos comparamos as amostras de tecidos em classificação a níveis de fumantes e verificamos em nossos resultados que existem semelhança nos dados de expressão entre as classificações, podemos comparar com os resultados de (GOLDMANN et al., 2021) que concatenou as classificações de fumantes em um único conjunto de indivíduos com câncer.

Com esses resultados verificamos a possibilidade de utilização da ferramenta rtranscriptograma para análise de resultados de atividade transcricional entre os conjuntos oriundos de microarranjo (GSE10072) e RNA-seq (GSE81089).

6 *Conclusões*

A utilização do *dashboard* com as funções de ordenamento e modularidade permitiram a construção de sub-redes e visualização.

A criação de uma interface gráfica de usuário (GUI) integrando todas as ferramentas permitiu através de caixas de diálogos orientar o usuário sobre as etapas do *workflow* e possíveis mensagens de erro, além de ser mais amigável pela utilização de caixas de seleções para carregar os arquivos necessários e botões para ajustes nos parâmetros de execução dos processos. Foram minimizados os arquivos de entrada para a execução de todos os processos, sendo necessário somente o arquivo de interação dos nodos da rede biológica e os arquivos de expressão de microarranjo ou RNA-seq. As etapas de seleção de quantidade de comparações, controles e testes foram otimizadas para serem executadas diretamente do *dashboard*. Como arquivos de saída é possível salvar em variáveis da linguagem de programação R ou mesmo em arquivos de ambientes externos (.RData) os resultados de todas as etapas para uma análise mais analítica.

Ao realizarmos as análises com dados de expressão oriundos das séries GSE10072 (microarranjo) e GSE81089 (RNA-seq) obtivemos similaridades entre os resultados das comparações e conseguimos aumentar a acurácia em relação aos resultados publicados desses dados. Também foi possível analisar todo o organismo de uma forma mais ampla, sem descarte de genes, por utilizar como base a rede proteica como referência.

Referências Bibliográficas

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics*, APS, v. 74, n. 1, p. 47, 2002.

ALHARTHI, A. *Addressing barriers to big data*. Amsterdam: Elsevier, 2017. 285-292 p.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, v. 25, n. 1, p. 25–29, May 2000. Disponível em: <<http://dx.doi.org/10.1038/75556>>.

BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, Nature Publishing Group, v. 5, n. 2, p. 101–113, 2004.

BLAZOTTI, A. A. *Desenvolvimento de Ferramenta Computacional para integração de transcriptomas e redes biológicas: medidas de desempenho global*. Dissertação (Mestrado) — Instituto de Biociências de Botucatu, Universidade Estadual Paulista "Julio de Mesquita Filho", 2016.

BROWN, N. et al. Big data in drug discovery. *Progress in medicinal chemistry*, v. 57, p. 277–356, 2018. ISSN 0079-6468.

BRUGGEMAN, F. J.; WESTERHOFF, H. V. The nature of systems biology. *Trends in microbiology*, v. 15, p. 45–50, jan. 2007. ISSN 0966-842X.

CAMILO, E. et al. Galant: a cytoscape plugin for visualizing data as functional landscapes projected onto biological networks. *Bioinformatics*, Oxford University Press, v. 29, n. 19, p. 2505–2506, 2013.

CARBONE, F.; MONTECUCCO, F. Big data and data sharing: opportunities for the urgent challenges in cardiovascular disease. 2019. ISSN 0014-2972.

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

CHARITOU, T.; BRYAN, K.; LYNN, D. J. Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution*, BioMed Central, v. 48, n. 1, p. 27, 2016.

COSTA, F. F. Big data in biomedicine. v. 19, p. 433–440, 2014. ISSN 1359-6446.

CZAUDEANA, T.; SCHREIBER, F. Information visualization for biological data. *Methods in molecular biology (Clifton, N.J.)*, v. 1526, p. 403–415, 2017. ISSN 1940-6029.

DALMA-WEISZHAUSZ, D. D. et al. The affymetrix genechip platform: an overview. *Methods in enzymology*, v. 410, p. 3–28, 2006. ISSN 0076-6879.

- DEITEL, H. M.; DEITEL, P. J. *C++ como programar*. [S.l.]: Prentice Hall, 1999.
- DUDOIT, S. et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, JSTOR, p. 111–139, 2002.
- EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, Oxford Univ Press, v. 30, n. 1, p. 207–210, 2002.
- ERDŐS, P.; RÉNYI, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, v. 5, n. 1, p. 17–60, 1960.
- FABREGAT, A. et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, v. 46, n. D1, p. D649–D655, 2018.
- FRASER, H. B. Modularity and evolutionary constraint on proteins. *Nat Genet*, London, UK, v. 37, n. 4, p. 351–352, Apr 2005. Disponível em: <<http://dx.doi.org/10.1038/ng1530>>.
- GOLDMANN, T. et al. Pd-l1 amplification is associated with an immune cell rich phenotype in squamous cell cancer of the lung. *Cancer immunology, immunotherapy : CII*, v. 70, p. 2577–2587, set. 2021. ISSN 1432-0851.
- GONZALEZ-ANGULO, A. M.; HENNESSY, B. T.; MILLS, G. B. Future of personalized medicine in oncology: a systems biology approach. *Journal of Clinical Oncology*, American Society of Clinical Oncology, v. 28, n. 16, p. 2777–2783, 2010.
- GOODMAN, D. *Dynamic HTML: The Definitive Reference: A Comprehensive Resource for HTML, CSS, DOM & JavaScript*. [S.l.]: "O'Reilly Media, Inc.", 2002.
- HAN, Y. et al. Advanced applications of RNA sequencing and challenges. *Bioinformatics and biology insights*, v. 9, p. 29–46, 2015. ISSN 1177-9322.
- HEATHER, J. M.; CHAIN, B. The sequence of sequence: The history of sequencing DNA. *Genomics*, v. 107, n. 1, p. 1–8, January 2016.
- HRDLICKOVA, R.; TOLOUE, M.; TIAN, B. RNA-seq methods for transcriptome analysis. *Wiley interdisciplinary reviews. RNA*, v. 8, jan. 2017. ISSN 1757-7012.
- HUANG, Y.-W.; HUANG, T. H.-M.; WANG, L.-S. Profiling DNA methylomes from microarray to genome-scale sequencing. *Technology in cancer research & treatment*, v. 9, p. 139–147, abr. 2010. ISSN 1533-0338.
- HUBER, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, Nature Publishing Group, v. 12, n. 2, p. 115, 2015.
- HUGHES, T. R. et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature biotechnology*, v. 19, p. 342–347, abr. 2001. ISSN 1087-0156.
- HWANG, B.; LEE, J. H.; BANG, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, Nature Publishing Group, v. 50, n. 8, p. 1–14, 2018.

- IDEKER, T.; GALITSKI, T.; HOOD, L. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 2, n. 1, p. 343–372, 2001.
- JALALI, S. et al. Systematic transcriptome wide analysis of lncrna-mirna interactions. *PloS one*, Public Library of Science, v. 8, n. 2, p. e53823, 2013.
- KADADI, A. et al. Challenges of data integration and interoperability in big data. In: . Washington, DC, USA: IEEE, 2014. p. 38–40. ISBN 978-1-4799-5666-1.
- KANEHISA, M. et al. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, v. 40, n. Database issue, p. D109–D114, Jan 2012. Disponível em: <<http://dx.doi.org/10.1093/nar/gkr988>>.
- KHOURY, M. J.; IOANNIDIS, J. P. A. Big data meets public health. v. 346, p. 1054–1055, 2014. ISSN 0036-8075.
- KUHN, M. et al. Stitch 4: integration of protein–chemical interactions with user data. *Nucleic acids research*, Oxford Univ Press, p. gkt1207, 2013.
- LANDI, M. T. et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS one*, Public Library of Science, v. 3, n. 2, p. e1651, 2008.
- LEINONEN, R.; SUGAWARA, H.; SHUMWAY, M. The sequence read archive. *Nucleic acids research*, Oxford Univ Press, p. gkq1019, 2010.
- LIU, J. et al. A prognostic 14-gene expression signature for lung adenocarcinoma: A study based on tcga data mining. *Oxidative medicine and cellular longevity*, v. 2020, p. 8847226, 2020. ISSN 1942-0994.
- LIU, L. et al. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, v. 2012, p. 251364, 2012. ISSN 1110-7251.
- LIU, Y. et al. Systematic analysis of mrna expression profiles in nscic cell lines to screen metastasis-related genes. *Molecular medicine reports*, v. 14, p. 5093–5103, dez. 2016. ISSN 1791-3004.
- LUN, A. T. L.; MCCARTHY, D. J.; MARIONI, J. C. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, v. 5, p. 2122, 2016. ISSN 2046-1402.
- METZKER, M. L. Sequencing technologies - the next generation. *Nature reviews genetics*, Nature Publishing Group, v. 11, n. 1, p. 31–46, 2010.
- MILLER, M. B.; TANG, Y.-W. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, Am Soc Microbiol, v. 22, n. 4, p. 611–633, 2009.
- MOLAN, A. L.; RYBARCZYK-FILHO, J. L. Desenvolvimento e compara  o de algoritmos para a organiza  o hier rquica de redes. *Anais do X Congresso de F sica Aplicada   Medicina*, v. 1, n. 117-121, 2014.

MOREIRA, L. M. *Ciencias genômicas : fundamentos e aplicacoes*. Sociedade Brasileira de Genetica, 2015. Disponível em: <<https://drive.google.com/file/d-/0Bwb0CldyHPwV3IxRGdyYklIdGc/view>>.

NELSON, D. L.; COX, M. M. *Lehninger Principles of Biochemistry, Fourth Edition*. Fourth edition. New York, USA: [s.n.], 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <<https://www.R-project.org/>>.

RISTEVSKI, B.; CHEN, M. Big data analytics in medicine and healthcare. *Journal of integrative bioinformatics*, De Gruyter, v. 15, n. 3, 2018.

RYBARCZYK-FILHO, J. L. et al. Towards a genome-wide transcriptogram: the *saccharomyces cerevisiae* case. *Nucleic acids research*, Oxford Univ Press, v. 39, n. 8, p. 3005–3016, 2011.

SAGIROGLU, S.; SINANC, D. Big data: A review. In: . San Diego, CA, USA: IEEE, 2013. p. 42–47. ISBN 978-1-4673-6402-7.

SANCHEZ-PINTO, L. N.; LUO, Y.; CHURPEK, M. M. Big data and data science in critical care. *Chest*, v. 154, p. 1239–1248, nov. 2018. ISSN 1931-3543.

SANGER, F.; NICKLEN, S.; COULSON, A. R. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 74, n. 12, p. 5463–5467, 1977.

SHAHZAD, K.; LOOR, J. J. Application of top-down and bottom-up systems approaches in ruminant physiology and metabolism. *Current genomics*, v. 13, p. 379–394, ago. 2012. ISSN 1389-2029.

SONIS, S.; VILLA, A. *Translational Systems Medicine and Oral Disease*. [S.l.]: Academic Press, 2019.

STRING. *STRING*. 2017. <http://string-db.org/>. [Online; accessed 17-Dec-2019].

SU, H. et al. Identification of potential biomarkers of lung adenocarcinoma brain metastases via microarray analysis of cDNA expression profiles. *Oncology letters*, v. 17, p. 2228–2236, fev. 2019. ISSN 1792-1074.

SZCZEŚNIAK, M. W.; MAKALOWSKA, I. lncRNA-rna interactions across the human transcriptome. *PloS one*, Public Library of Science, v. 11, n. 3, p. e0150353, 2016.

SZKLARCZYK, D. et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, Oxford Univ Press, p. gku1003, 2014.

TANG, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nature methods*, v. 6, p. 377–382, maio 2009. ISSN 1548-7105.

TREVINO, V.; FALCIANI, F.; BARRERA-SALDANA, H. A. Dna microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular medicine (Cambridge, Mass.)*, v. 13, p. 527–541, 2007. ISSN 1076-1551.

TSENG, G. C.; GHOSH, D.; FEINGOLD, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, v. 40, p. 3785–3799, maio 2012. ISSN 1362-4962.

VERLI, H. Bioinformática: da biologia à flexibilidade molecular. Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

VINOGRADOV, A. E. Modularity of cellular networks shows general center-periphery polarization. *Bioinformatics*, Oxford, UK, v. 24, n. 24, p. 2814–2817, Dec 2008. Disponível em: <<http://dx.doi.org/10.1093/bioinformatics/btn555>>.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, v. 10, p. 57–63, jan. 2009. ISSN 1471-0064.

WINNICK, E. Dna sequencing industry sets its sights on the future: new technologies may duke it out with upgrades of the current method. *The Scientist*, Scientist Inc., v. 18, n. 18, p. 44–48, 2004.