UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Allan Mesquita da Costa

Métodos computacionais para a análise de dados de expressão gênica provenientes de uma análise de *microarray* utilizada para teste farmacológico.

Recife

2023

Allan Mesquita da Costa

Métodos computacionais para a análise de dados de expressão gênica provenientes de uma análise de *microarray* utilizada para teste farmacológico.

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Jeane Cecília Bezerra de Melo

Coorientador: Luciana Amaral de Mascena Costa

Recife

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

C838m COSTA, ALLAN MESQUITA
Métodos computacionais para a análise de dados de expressão gênica provenientes de uma análise de microarray utilizada para teste farmacológico. / ALLAN MESQUITA COSTA. - 2023.
50 f. : il.

Orientadora: Jeane Cecilia Bezerra de Melo.
Coorientadora: Luciana Amaral de Mascena Costa.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2023.

1. Microarray. 2. Biologia Computacional. 3. Clusterização. 4. Análise de Dados. I. Melo, Jeane Cecilia Bezerra de, orient. II. Costa, Luciana Amaral de Mascena, coorient. III. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Allan Mesquita da Costa às 9 horas do dia 28 de abril de 2023, no link https://meet.google.com/ezg-rjih-ioh, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado "Métodos computacionais para a análise de dados de expressão gênica provenientes de uma análise de Microarray, utilizando para teste farmacológico", orientado por Jeane Cecília Bezerra de Melo e aprovado pela seguinte banca examinadora:

_____

Jeane Cecília Bezerra de Melho
DC/UFRPE

_____

Manoel Adrião Gomes Filho
DMFA/UFRPE

Dedico também aos meus pais, que sempre foram exemplos de dedicação, esforço e amor incondicional. Obrigado por me ensinarem a ser perseverante e por me apoiarem em todas as decisões que tomei ao longo da minha vida.

Não poderia deixar de dedicar também aos meus irmãos, que sempre foram meus amigos, companheiros de jornada e me ensinaram a importância da união e da amizade verdadeira.

A todos vocês, que fazem parte da minha vida e me ajudaram em todos os momentos, dedico este trabalho e espero que nossa união e amor só aumentem a cada dia. Obrigado por estarem sempre ao meu lado!*Dedico este trabalho de conclusão de curso à minha amada esposa, que sempre esteve ao meu lado me apoiando, incentivando e mostrando que juntos somos capazes de enfrentar qualquer desafio.*

Dedico também aos meus pais, que sempre foram exemplos de dedicação, esforço e amor incondicional. Obrigado por me ensinarem a ser perseverante e por me apoiarem em todas as decisões que tomei ao longo da minha vida.

Não poderia deixar de dedicar também aos meus irmãos, que sempre foram meus amigos, companheiros de jornada e me ensinaram a importância da união e da amizade verdadeira.

A todos vocês, que fazem parte da minha vida e me ajudaram em todos os momentos, dedico este trabalho e espero que nossa união e amor só aumentem a cada dia. Obrigado por estarem sempre ao meu lado!

Agradecimentos

À minha amada esposa Luciana Costa, minha gratidão, por seu constante apoio, encorajamento e compreensão durante todo o processo de elaboração deste trabalho de conclusão de curso. Sua dedicação e carinho me ajudaram a superar momentos de dificuldade e a manter a motivação para buscar sempre o melhor resultado possível. Obrigado por estar sempre ao meu lado, por acreditar em mim e por ser minha parceira em todas as jornadas da vida. Sem a sua presença, certamente não teria sido possível chegar até aqui. Com toda a minha gratidão e amor.

Aos meus pais Nilton e Maria das Graças, gratidão e reconhecimento por todo o apoio, incentivo e amor que vocês têm me dado ao longo de minha jornada acadêmica. Sei que sem a ajuda de vocês, eu não teria conseguido chegar até aqui. Muito obrigado por cada momento de encorajamento, paciência e colaboração durante a pesquisa, escrita e defesa deste trabalho. Seu apoio constante foi fundamental para que eu pudesse superar os desafios e obstáculos que surgiram ao longo do caminho. Espero que esta conclusão seja também um motivo de orgulho para vocês e que possa retribuir de alguma forma todo o suporte que recebi.

À minha orientadora, minha profunda gratidão por todo o apoio, aconselhamento e encorajamento que você me deu durante todo o processo de elaboração do meu trabalho de conclusão de curso (TCC). Sua orientação foi essencial para o sucesso desse projeto. Desde o início, você esteve presente, disposta a responder todas as minhas dúvidas e fornecer orientações valiosas sobre como organizar e desenvolver meu trabalho. Sem você, eu certamente teria me sentido mais perdido e ansioso em relação a essa tarefa.

“Fortuna favet audaci.”
(Autor desconhecido)

Resumo

O advento do Projeto Genoma Humano (PGH), finalizado em outubro de 2003, impulsionou o desenvolvimento de técnicas para obtenção e análise de dados biológicos. A necessidade de gerenciar o grande volume de dados do genoma digital foi um fator determinante no crescimento de uma área de conhecimento multidisciplinar, a Biologia Computacional. Nas duas décadas subsequentes à finalização do PGH, genomas de diferentes organismos foram obtidos. Em relação aos mamíferos, projetos tais como o 1000 *Genomes Project* e o *Cancer Genome* Atlas (TCGA) ilustraram o avanço de conhecimento na análise de dados complexos. Dentre as técnicas mais recentes, destacamos os *Microarrays*. Estes fornecem uma quantidade significativa de dados em um único experimento, permitindo a comparação de genomas completos. A análise de dados de *Microarray* é relativamente complexa e demanda por protocolos que tornem esta análise mais simples, produzindo informações mais compreensíveis. O presente estudo compreende a utilização de métodos computacionais para analisar dados de expressão de gênica obtidos de um experimento de *Microarray* utilizado para teste farmacológico relativos ao câncer de mama. Para o processamento dos dados brutos, obtidos de uma planilha contendo mais de 3216 genes resultantes de uma análise de *Microarray*, foi desenvolvido um *script* visando facilitar a extração de informação a partir destes dados e posterior seleção dos genes de interesse. O programa possibilitou a busca dos genes envolvidos nos processos de morte celular (apoptose, necrose e autofagia), os quais são fatores determinantes na análise de sucesso do fármaco testado. Para a categorização dos genes envolvidos na cascata de morte apoptótica, necrótica e autofágica, foram construídos *heatmaps* a partir dos valores dos níveis de expressão chamados de *fold-change* diferença da expressão gênica para os valores antes e depois do tratamento das células cancerígenas com o composto mesoiônico), utilizando técnicas de clusterização *k-means* e clusterização hierárquica disponibilizadas no programa *Heatmapper*. Como resultados do foi desenvolvido um *script* no programa R que resultou na separação de 20 genes envolvidos na cascata de morte de morte apoptótica, seis envolvidos na morte autofágica e 7 envolvidos na morte necrótica, além do desenvolvimento de 3 *Heatmaps*, contribuindo para a análise biológica dos dados, além de tornar mais acessível o processamento dos dados de *Microarray*.

Palavras-chave: *Microarray*, Biologia Computacional, Clusterização e Análise de Dados.

Abstract

The advent of the Human Genome Project (HGP), completed in October 2003, propelled the development of techniques for obtaining and analyzing biological data. The need to manage the vast amount of digital genome data was a decisive factor in the growth of a multidisciplinary area of knowledge, Computational Biology. In the two decades following the completion of the HGP, genomes of different organisms were obtained. Regarding mammals, projects such as the 1000 Genomes Project and the Cancer Genome Atlas (TCGA) illustrated the advancement of knowledge in the analysis of complex data. Among the newest techniques, we highlight Microarrays. They provide a significant amount of data in a single experiment, allowing the comparison of complete genomes. The analysis of Microarray data is relatively complex and requires protocols that make this analysis simpler, producing more understandable information. The present study involves the use of computational methods to analyze gene expression data obtained from a Microarray experiment used for pharmacological testing related to breast cancer. To process the raw data, obtained from a spreadsheet containing more than 3216 genes resulting from a Microarray analysis, a script was developed to facilitate the extraction of information from this data and subsequent selection of genes of interest. The program allowed the search for genes involved in the processes of cell death (apoptosis, necrosis, and autophagy), which are determining factors in the success analysis of the tested drug. To categorize the genes involved in the apoptotic, necrotic, and autophagic death cascade, heatmaps were constructed from fold-change values (difference in gene expression for values before and after treatment of cancerous cells with the mesoionic compound), using k-means clustering and hierarchical clustering techniques provided in the Heatmapper program. Results of the study include the development of a script in the R program that resulted in the separation of 20 genes involved in the apoptotic death cascade, six involved in the autophagic death, and seven involved in the necrotic death cascade, as well as the development of three heatmaps, contributing to the biological analysis of data, in addition to making Microarray data processing more accessible.

Keywords: Microarray, Computational biology, Clustering and Data Analytics.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

cDNA	DNA complementar
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DNA	Ácido Desoxirribonucleico
PCR	Reação em Cadeia da Polimerase
PGH	Projeto Genoma Humano
MeV	Visualizador de Múltiplos Experimentos
MIH2.4	Composto mesoiônico
mRNA	RNA mensageiro
NCBI	National Center for Biotechnology Information
RNA	Ácido Ribonucleico
SOM	Mapas de Auto-organização
TIGR	The Institute for Genomic Research

Sumário

1 Introdução

Biologia computacional uma área multidisciplinar, abrangente e complexa, na qual inúmeros modelos e ferramentas vêm sendo desenvolvidas ao longo dos anos, possibilitando assim, a análise de grandes volumes de dados, bem como a obteção e tratamento de informação de forma mais ágil. Esses dados podem ser, por exemplo, resultantes de experimentos na área de Biologia Molecular. O Projeto Genoma Humano, além de impulsionar a Biologia Computacional como área de estudos, trouxe uma nova maneira de desenvolvermos pesquisas em Biologia (BIRNEY, 2021), na qual faz-se necessário o uso de métodos advindos de outras áreas, tais como Ciência da Computação, Matemática e Estatística, visando a compreensão de algumas informações biológicas substanciais (HOOD, 2003). Essas mudanças nos levam a uma nova abordagem da biologia, denominada Biologia de Sistemas, a qual estuda as inter-relações de todos os elementos em um sistema em vez de estudá-los separadamente, como eventos isolados (HOOD, 2003), (AZAD; SHULAEV, 2019), (JOSHI et al., 2021), (PELLENZ; CRISPIM; ASSMANN, 2022).

Assim, a Biologia computacional desponta em aplicações que envolvem, por exemplo, a ordenação dos resultados adivindos das iniciativas de sequenciamento de genes, produzindo um volume cada vez maior de dados sobre sequências de DNA e seus produtos proteicos, e mesmo a modelagem de soluções viáveis para problemas mais complexos (ZHANG et al., 2022), (VIEIRA; LAUBENBACHER, 2022), (PRIYAMVADA et al., 2022). Desse modo, os biologistas moleculares têm criado implementações para métodos estatísticos capazes de analisar grandes quantidades de dados biológicos, de inferir funções dos genes e de estabelecer relações estruturais entre genes e proteínas (ARAÚJO et al., 2008).

É sabido que a união entre a informatização e a moderna biologia molecular contribuiu para o surgimento dessa nova área do conhecimento. Entretanto, o advento da bioinformática teve que esperar algum tempo para realmente ser incorporada ao contidiano de pesquisas nessa área multidiscplinar. Apesar da estrutura do DNA ter sido desvendada em 1953 pelos ingleses Watson e Crick (GRIFFITHS et al., 2006), foi preciso esperar até a metade da década de 1980 para que fosse desenvolvida uma lente de aumento suficientemente boa (uma máquina automatizada) que permitisse a leitura em grandes quantidades do código genético contido nas biomoléculas (ARAÚJO et al., 2008). Nesse mesmo período, a Ciência da Computação também apresentou avanços significativos, com o advento de computadores com capacidade de armazenamento cada vez maiores, aumento significativo da velocidade de processamento, a custos reduzidos.

Um outro avanço promovido pelo Projeto Genoma Humano foi o desenvolvimento de sequenciadores cada vez melhores. Portanto, os sequenciadores de última geração refinaram o estudo de doenças genéticas humanas, através do sequenciamento do exoma (região codificadora do genoma), detectando variantes causadores de doenças e descobrindo alvos genéticos (PERVEZ et al., 2022), (GARRIDO-CARDENAS et al., 2017), (BAO et al., 2014). Assim, o **Sequenciamento Completo do Exoma** consiste na análise dos éxons do DNA, que correspondem às regiões codificantes do mRNA. A análise de exoma busca detectar os genes (genômica), mRNA (transcriptômica), proteínas (proteômica) e metabólitos (metabolômica) em uma amostra biológica específica.

Tecnologias de alta produtividade -omics", como *Microarray*, sequenciamento, e Reação em Cadeia da Polimerase em tempo real (PCR), têm sido amplamente aplicados em análises biológicas e pesquisas biomédicas a mais de uma década (VIJAYAKUMAR et al., 2018), (FERTIG; SLEBOS; CHUNG, 2012), tendo como objetivos identificar e quantificar todas as biomoléculas de um tipo específico (DNA, RNA, proteína e metabólito) em uma dada amostra biológica. Esse tipo de estudo dependente desse conjunto amplo de dados, denominado multi-omic, permite interpretar a complexidade molecular e variações em diversos níveis, como genoma, epigenoma, transcriptoma, proteoma e metaboloma (SUBRAMANIAN et al., 2020).

1.1 Motivação

O presente trabalho reporta estudos relacionados a uma tecnologia recente nos estudos genômicos, os já citados, *Microarrays*. A tecnologia de *Microarray* é uma técnica da biologia molecular que permite a busca sistemática de genes humanos, animais e plantas que possam estar envolvidos em diversas patologias, com o objetivo de dignóstico, prognóstico e possíveis alvos terapêuticos, gerando assim um grande volume de dados que precisam ser analisados para serem interpretados, demandando conhecimentos de ciência da computação, nem sempre dominados pelos pesquisadores biólogos e da área de saúde (AGAPITO; ARBITRIO, 2022), motivando o desenvolvimento de ferramentas que permitam simplicar esse tipo de análise por profissionais que não possuem formação na área de ciências exatas.

Para desenvolver a pesquisa, buscamos um conjunto de dados inédito, relacionado a experimentos de *Microarray*, especificamente, uma análise de teste farmacológico relativos ao câncer de mama (COSTA et al., 2020). O estudo foi conduzido a partir da pergunta de pesquisa exibida a seguir.

1.2 Pergunta de Pesquisa

Quais métodos computacionais podem auxiliar na análise de dados de expressão gênica, obtidos de um experimento de *Microarray* utilizado para a análise de teste farmacológico relativos ao câncer de mama?

1.3 Justificativa

Propor um conjunto de métodos computacionais para auxiliar na análise de *Microarray* é uma problema atual e relevante, uma vez que permite observar mais rapidamente os resultados advindos do experimento (AGAPITO; ARBITRIO, 2022). Ao consideramos o conjunto de dados inéditos utilizados nesse estudo (COSTA et al., 2020), observamos que a natureza da pesquisa para esse experimento possui suas especificidades, as quais delimitaram a pesquisa aqui apresentada, que foi conduzida de forma a atender a demanda específica, mas também, permitir a generalização para utilização em outros tipos de experimentos.

1.4 Estrutura do Trabalho

O presente Trabalho de Conclusão de Curso está organizado da seguinte forma: O Capítulo 1 traz uma breve apresentação da área de estudo, apresenta o tema desenvolvido, a motivação para estudá-lo, os dados utilizados na pesquisa, que delimitaram nossa pergunta de pesquisa, a justificativa e como este documento está organizado. Uma revisão exploratória sobre trabalhos científicos relacionados ao tema de estudo compõe o Capítulo 2, que também versa sobre alguns conceitos necessários à compreensão desse estudo. Os objetivos gerais e específicos dão descritos no Capítulo 3. O Capítulo 4, por sua vez, contém a descrição de materiais e metódos utilizados durante o desenvolvimento do presente estudo. Resultados e discussões sobre estes são apresentados no Capítulo 5. As conclusões e trabalhos futuros compõem o Capítulo 6.

2 Revisão bibliográfica

2.1 Biologia Computacional

Desde a descoberta do DNA, na década de 50, como a molécula que armazena a informação genética e elucidação da sua estrutura molecular, esta molécula passou a ser o principal foco de estudo da biologia molecular. O processo de sequenciamento iniciou-se na década de 70 através de um processo manual e rudimentar sendo sequenciada de 10 a 100 bases por vez (FARIAS; CHACON; SILVA, 2012). Após o projeto genoma humano, resultado de um experimento bem-sucedido e amplamente conhecido, surge a era genômica e o consequente desenvolvimento das ferramentas de Bioinformática como novo ramo científico, os quais antes seriam inviáveis sem o desenvolvimento de tecnologias para o sequenciamento e análise dos dados obtidos (ARAÚJO et al., 2008).

Pauline Hogeweg em 1979 utilizou pela primeira vez o termo “Bioinformática”, para estudos de processos de informática em estudos de biologia sistemacional (ALVES, 2013). A literatura diverge em relação a definição do termo Bioinformática, mas de modo geral, a Bioinformática envolve a aplicação de Tecnologias de Informação e de Comunicação (TIC) nas análises de qualquer área da Biologia. De maneira mais específica, a Bioinformática é a aplicação de informática aos experimentos de Biologia Molecular, ou mais especificamente no manejo da grande quantidade de dados proveniente por exemplo do sequenciamento de DNA, RNA e Genômica (ALVES, 2013).

É possível entender a biologia fazendo uma analogia a ferramentas digitais, por exemplo, quando pensamos que a célula possui um núcleo digital de informações (genoma) a partir do qual o envelhecimento é possível e conhecendo apenas os processos biológicos iniciais. O desafio fundamental seria, entender como o envelhecimento acontece, a partir dos processos biológicos iniciais, entender como as células obtêm informações desse núcleo genômico digital e como elas conseguem converter essas informações, de uma única célula (o ovo fertilizado) em um organismo adulto (para humanos, 10^{14} células) com muitos tipos diferentes de células. O ponto importante é o código do genoma digital de todos os organismos vivos, o “código-fonte” para o desenvolvimento, que foi finalmente conhecido com a conclusão do projeto genoma humano (HOOD, 2003).

Com o advento dos sequenciadores de última geração e o investimento no projeto genoma humano, ocorreu um grande avanço no estudo de doenças genéticas humanas, através do sequenciamento do exoma (região codificadora do genoma) a um nível profundo, detectando variantes causadores de doenças e descobrindo alvos genéticos.

(BAO et al., 2014). No período evolutivo, os seres humanos herdaram um modo particular para formação das proteínas. Este modo consiste em duas etapas principais, na primeira as bases de DNA que formam os genes são transcritas em uma molécula RNA (processo de transcrição), nessa molécula existirão regiões que codificarão proteínas, chamadas de exóns e regiões não codificadoras chamadas de íntrons. Após a molécula de RNA ser formada, ela necessita passar por um processo conhecido como *splicing* no qual essas regiões que não codificarão (íntrons) são removidas, dando origem a moléculas de RNA maduras. O RNA maduro, conhecido como RNA mensageiro (mRNA), contém apenas as bases que codificarão diretamente as proteínas (ROEHRS et al., 2009). Ao todo, o somatório dos éxons representa aproximadamente 45 milhões de pares de bases. O Sequenciamento Completo do Exoma consiste na análise dos éxons do DNA, que correspondem às regiões codificantes do mRNA.

A partir da análise do exoma, tecnologias conhecidas como 'Omic' adotam uma visão holística das moléculas que constituem uma célula, tecido ou organismo. Eles são voltados principalmente para a detecção universal de genes (genômica), mRNA (transcriptômica), proteínas (proteômica) e metabólitos (metabolômica) em uma amostra biológica específica, de forma não direcionada e não tendenciosa. Essa abordagem também pode ser referida como biologia de alta dimensão; a integração dessas técnicas é chamada de biologia dos sistemas (KELL, 2007) (WESTERHOFF; PALSSON, 2004). O aspecto básico dessas abordagens é que um sistema complexo pode ser bem compreendido, se for observado como um todo. A Biologia de sistemas e "ômicas" diferem dos estudos tradicionais, pois estes são amplamente baseados em hipóteses ou ideias reducionistas. Em contraste, experimentos de biologia de sistemas geram hipóteses, usando abordagens holísticas onde nenhuma hipótese é conhecida ou prescrita, mas todos os dados são adquiridos e analisados para definir uma hipótese, que poderá ser testada posteriormente (KELL, 2007).

Tecnologias de alta produtividade -omics", como *microarray*, sequenciamento, e Reação em Cadeia da Polimerase em tempo real (PCR), têm sido amplamente aplicados em análises biológicas e pesquisas biomédicas a mais de uma década (FERTIG; SLEBOS; CHUNG, 2012). Sendo essas tecnologias utilizadas para identificar e quantificar todas as biomoléculas de um tipo específico (DNA, RNA, proteína e metabólito) em uma dada amostra biológica. Esse tipo de estudo, fez com que a biologia se tornasse cada vez mais dependente dos dados gerados por eles, que juntos são chamados de dados "multi-omics", para assim, conseguirem interpretar a complexidade molecular e variações em diversos níveis, como genoma, epigenoma, transcriptoma, proteoma e metaboloma (SUBRAMANIAN et al., 2020).

A bioinformática baseia-se na elucidação de dados biológicos brutos provenientes de por exemplo, de experimentos da biologia molecular, desenvolvendo programas

computacionais que permitam reconhecer sequências de genes; analisar experimentos de expressão gênica, predizer a configuração tridimensional de proteínas; identificar inibidores de enzimas; organizar e relacionar informação biológica; agrupar conjuntos de proteínas homólogas, simular estruturas tridimensionais de proteínas-alvos, entre outros (ARAÚJO et al., 2008). Fundamentada nas aplicações da informática para a biologia, pode-se dizer que a Bioinformática é uma ciência interdisciplinar pela fusão de conhecimentos biológicos, informáticos e matemáticos para a análise de informações de sequências genômicas e para a predição de aspectos estruturais e funcionais de biomoléculas, incluindo armazenamento, gerenciamento, análise, recuperação e visualização de dados biológicos, muitas vezes em sistemas modelos, transformando dados em informações e conhecimento (LESK, 2008).

Os principais recursos da bioinformática são os programas de computadores e os bancos de dados disponíveis na rede, ação fundamental para a análise de sequências de DNA e proteínas. Esta ferramenta é capaz de promover o aumento da velocidade na análise de sequências de DNA de diferentes fontes, na comparação de variabilidades e na previsão de resultados de análises (ARAÚJO et al., 2008).

2.2 Ferramentas da biologia computacional para análise de *microarray*

Dentre os métodos de análise dos dados de microarranjos, os mais utilizados são a construção de agrupamentos, seja para genes ou para amostras, a busca de genes diferencialmente expressos e a busca por grupos de genes capazes de distinguir tipos biológicos diferentes (análise de classificação ou discriminatória). Na análise de dados de expressão gênica, o agrupamento é visto como um método importante, pois ele viabiliza a detecção de grupos de genes que exibem padrões de expressão similares (co-expressos ou co-regulados) ou que mostram expressão diferencial (GIACHETTO, 2010). Para os processos celulares, é sabido que genes contidos em uma via particular ou que respondem a estímulos externos comuns podem ser co-regulados e, consequentemente, mostrar padrões similares de expressão (QUACKENBUSH, 2001). A formação de agrupamentos que se originam de dados numéricos observados experimentalmente encontra-se em uma área da estatística chamada de análise multivariada, e baseia-se em um método para agrupar os dados de acordo com uma medida de similaridade (GIACHETTO, 2010).

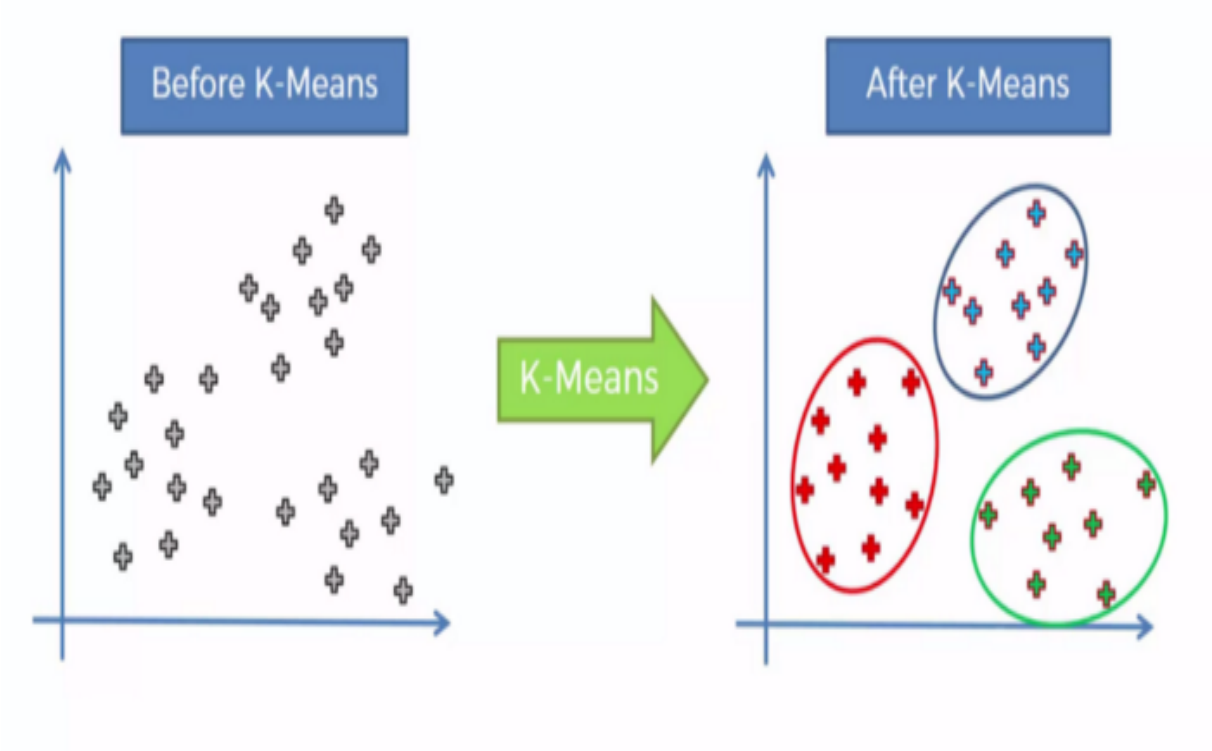
Baseado na análise dos microarranjos, a análise de agrupamentos pode ser compreendida como o processo de reunir elementos similares entre si, sendo que esses elementos podem ser os genes ou as amostras biológicas estudadas (ESTEVES, 2007). O agrupamento baseia-se em medidas de similaridade, ou métricas, que nada mais são

do que fórmulas matemáticas que calculam um número positivo a partir de dois pontos do seu espaço de elementos, no caso representadas pelo nível de expressão dos genes. Atualmente, a distância euclidiana e a medida de correlação têm sido métricas bastante utilizadas em análises de agrupamentos de dados de expressão gênica. Dentre os diferentes algoritmos existentes, e há uma série deles, pode-se citar o agrupamento hierárquico, o *K-means*, e o *Self-Organizing Maps* (SOM), como os mais utilizados.

2.2.1 Agrupamento *K-means*

O agrupamento *K-means* tem sido usado com sucesso para analisar dados de *microarray*, essa técnica separa os dados de maneira semelhante a mapas auto-organizados. O *K-means* agrupa os dados de expressão gênica indicando o número de grupos desejado (essa é uma limitação, pois é necessário saber o número de grupos) o centróide de cada um desses grupos é calculado e o algoritmo analisa a distância ou similaridade desses centróides com todos os elementos a serem agrupados; cada objeto é então alocado ao grupo cujo centróide esteja mais próximo e, ao ser incluído, esse centróide é recalculado para representar esse novo objeto (GOLLUB; SHERLOCK, 2006).

Figura 1 – *K-means*.

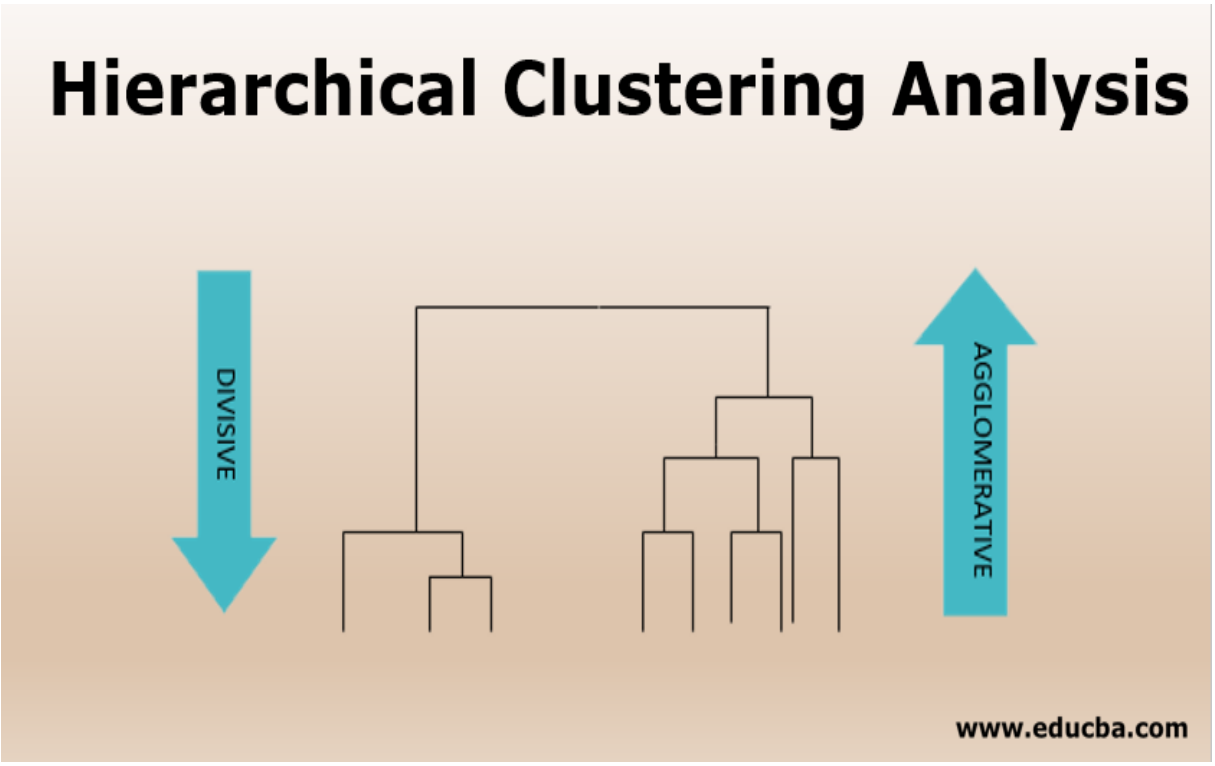


Fonte: http://exponentis.es/ejemplo-de-clustering-con-k-means-en-python

2.2.2 Agrupamento hierárquico

O agrupamento hierárquico é um procedimento aglomerativo que se baseia em agrupar as entidades de menor distância. Inicialmente, cada entidade representa um grupo. O par mais próximo é identificado constituindo um novo grupo. As distâncias entre esse novo grupo e as entidades restantes são calculadas, até que todos os elementos tenham sido agrupados (GIACHETTO, 2010). Baseado no agrupamento hierárquico existe a técnica de *clustering* na qual é classificada como sendo do tipo não supervisionada,fundamentada no princípio de que o algoritmo computacional é capaz de identificar por si só as classes dentro de um conjunto de dados (GONÇALVES et al., 2008). Embora exista uma grande quantidade de diferentes métodos de agrupamentos na área de reconhecimento de padrões (XU et al., 2005), a maioria dos *softwares* ou sistemas computacionais voltados para o processamento digital de imagens de sensoriamento remoto realiza a classificação não-supervisionada baseada em métodos de agrupamentos particionais, como o *K-means* e o ISODATA (BALL; HALL, 1967).

Figura 2 – *hierarquical clustering*.



Fonte: https://www.educba.com/hierarchical-clustering-analysis/

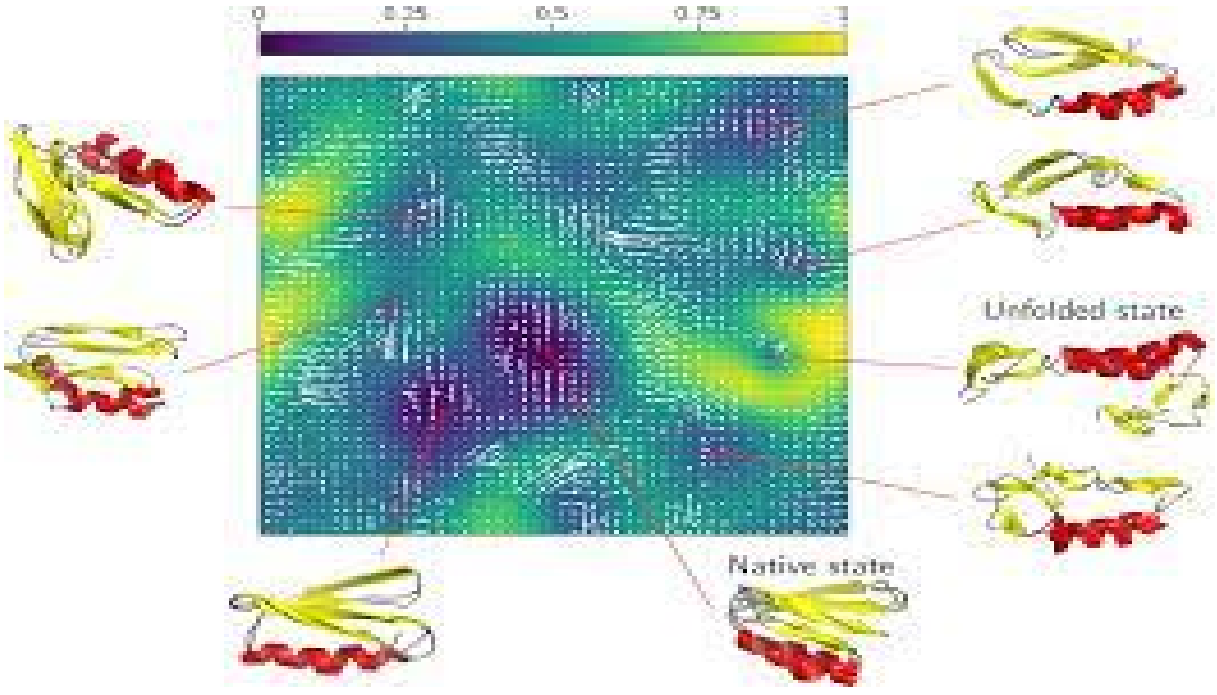
2.2.3 Agrupamento *Self-Organizing Maps* (SOM)

O algoritmo SOM é um dos mais comuns de rede neural artificial, frequentemente utilizado para tarefas de agrupamento e visualização. O agrupamento SOM é bastante similar ao método de *K-means*, onde um número pré-definido de grupos é especificado.

Contudo, nesse método, os grupos se relacionam por meio uma topologia espacial, geralmente arranjados em uma grade quadrada ou hexagonal, onde, inicialmente, os elementos são alocados aos seus grupos aleatoriamente. O algoritmo interativamente recalcula os centróides dos grupos baseado nos elementos de cada um, assim como aqueles elementos da vizinhança, e então realoca os elementos aos grupos. Uma vez que os grupos estão espacialmente relacionados, grupos vizinhos podem geralmente ser unidos ao final de uma interação, baseado num valor previamente estabelecido (GIACHETTO, 2010).

Existe também uma grande variedade de análises de microarranjo e sites gratuitos disponíveis, muitos dos quais implementam algumas formas de *clustering* (GOLLUB; SHERLOCK, 2006). Alguns como exemplos gratuitos e de código aberto de duas abordagens diferentes *software* de análise são o TIGR *Multiexperiment Viewer* (MeV) e o Linguagem de programação estatística R. A linguagem de programação R é uma linguagem de programação desenvolvida em parte para, e acompanhada por muitas funções para análise estatística. Como tal, é uma ferramenta poderosa para aqueles com algum conforto com programação de computador (e uma série de ferramentas de análise disponíveis também foram escritas em R). O BioCondutor projeto (GENTLEMAN et al., 2004) (REIMERS; CAREY, 2006) fornece um grande número de pacotes complementares para R especificamente destinado à pesquisa de microarray. De especial pertinência aqui, várias funções no R suportam agrupamento hierárquico e outras técnicas de análise não supervisionadas, bem como análises supervisionadas. A programação em R e BioConductor, são ferramentas poderosas para o biólogo computacional.

Figura 3 – *self-organizing map*.



Fonte: https://pypi.org/project/quicksom/

3 Objetivos

Uma vez apresentada uma revisão exploratória de estudos e conceitos relacionados a nossa pergunta de pesquisa apresentada no Capítulo 1:

- Quais métodos computacionais podem auxiliar na análise de dados de expressão gênica, obtidos de um experimento de *Microarray* utilizado para a análise de teste farmacológico relativos ao câncer de mama?

Estabelecemos no presente capítulo os objetivos, geral e específicos, definidos para o presente trabalho.

3.1 Objetivo Geral

O presente estudo compreende desenvolvimento de métodos computacionais para analisar dados de expressão de gênica obtidos de um experimento de *Microarray* utilizado para teste farmacológico relativos ao câncer de mama.

3.2 Objetivos Específicos

- Realizar estudos na área de Biologia Molecular visando a compreensão de Análise Genômica;
- Mapear as principais abordagens e métodos para estudo de análise de expressão gênica;
- Estudar as especificidades relativas as conjunto de dados utilizado;
- Aplicar técnicas de expressão gênica em conjuntos específicos de dados considerando técnicas de clusterização;
- Análise sistemática dos resultados obtidos que busquem promover interpretações biológicas para os diferentes problemas abordados.

A fim de alcançar tais ojetivos, foram estudados o conjunto de dados e suas especificidades, além de possíveis métodos de análise que possam ser aplicados nestes. Aqueles utilizados e desenvolvidos no presente estudo são apresentados no capítulo a seguir.

4 Metodologia

4.1 Apresentação

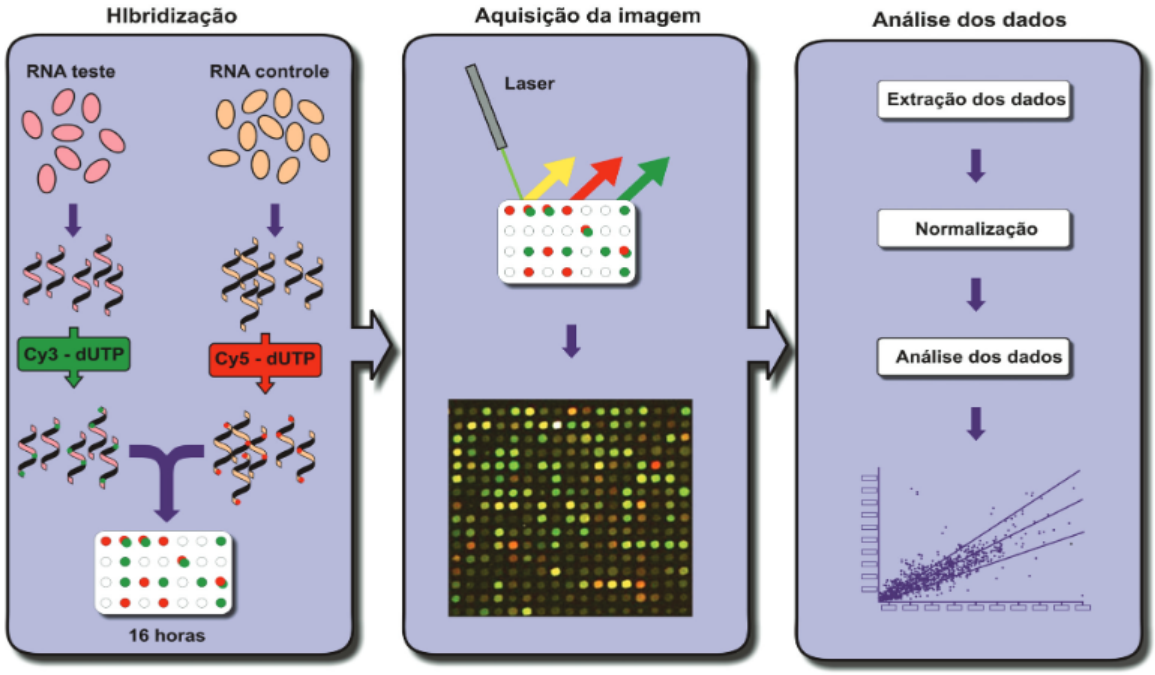
Neste capítulo apresentamos a metodologia utilizada para efetuar a análise de *microarray* tendo em vista a demanda especificada pelos biólogos que realizaram tal experimento de modo a realizar testes farmacológico relativos ao câncer de mama. O capítulo abrange desde as formas de análise de *microarray* até a descrição e apresentação do código dos *scripts* desenvolvidos para este fim.

4.2 Análise de *microarray*

O desenvolvimento do chip de DNA ou tecnologia de *microarray* é uma técnica que nos permite investigar a expressão de centenas ou milhares de genes em uma dada amostra, usando uma reação de hibridização. Durante o desenvolvimento dessa técnica, dois termos que são muito utilizados precisam ser melhor entendidos: sondas e alvos. As sondas são fragmentos gênicos conhecidos que estão imobilizados em uma matriz sólida, ou seja, as sequências correspondentes aos genes de interesse. Os alvos, por sua vez, consistem nos DNA complementares (cDNA), ou seja, o DNA sintetizado a partir de uma molécula de RNA mensageiro. Os cDNA estão marcados na solução e são provenientes das amostras biológicas (RUSSO; ZEGAR; GIORDANO, 2003).

Dessa forma, um *microarray* pode ser definido como uma coleção ordenada de sondas, cada sonda representando uma única espécie de ácido nucleico e correspondendo ao gene de interesse. A tecnologia é baseada na hibridização entre sequências de cDNA alvos, (Figura 4) derivadas das amostras de interesse que são marcadas com fluoróforos ou radioisótopos, com um *array* de várias sondas de DNA, que estão imobilizadas em uma matriz sólida. Essa matriz pode ser uma lâmina de vidro ou sílica ou uma membrana de nylon. Os fluorocromos mais utilizados são Cy3 (*Cyanine* 3) e Cy5 (*Cyanine* 5), os quais apresentam características distintas, o que permite que duas amostras diferentes possam ser co-hibridizadas na matriz sólida. O sinal de hibridização produzido em cada sonda corresponde ao nível de expressão do respectivo gene em determinada amostra no momento do estudo. Dessa forma, após a hibridização competitiva dos alvos marcados, os sinais são detectados, quantificados, integrados e normalizados com *softwares* específicos e refletem o perfil de transcrição gênica para cada amostra biológica (NIEMEYER; BLOHM, 1999).

Figura 4 – Esquema que demonstra o processo de hibridização dos dados até a análise dos dados.



Fonte: (COLOMBO; RAHAL, 2010)

Experimentos genômicos de alto rendimento, que envolvem o uso de *microarray* ou tecnologias de sequenciamento de última geração, são usados em muitos campos da biologia molecular, para descobrir genes envolvidos em processos biológicos específicos (AGAPITO; ARBITRIO, 2022). Para uma melhor compreensão dos resultados, análises de perspectiva global e ferramentas de visualização intuitiva são necessários para extrair informações relevantes (PEREZ-DIEZ; MORGUN; SHULZHENKO, 2007). Um importante aspecto na interpretação dos resultados de experimentos genômicos de alto rendimento, é a integração desses novos dados obtidos, a conhecimentos biológicos que já são conhecidos. Portanto, é importante que se tenha acesso a uma grande quantidade de conhecimentos biológicos, que foi previamente descrito, e que existam ferramentas que façam a integração de novos dados com os dados existentes (PEREZ-DIEZ; MORGUN; SHULZHENKO, 2007). Como exemplo pode ser citado o Biomart (SMEDLEY et al., 2009) é um sistema de integração de dados robusto, para consulta de dados em grande escala, usados para fornecer acesso fácil a uma série de bancos de dados biológicos importantes, como Ensembl, portal de dados ICGC, ArrayExpress, COSMIC entre outros (http://www.biomart.org) e o KEGG é um recurso de banco de dados para a compreensão de funções de alto nível e utilidades do sistema biológico, como a célula, o organismo e o ecossistema, a partir de informações em nível molecular, especialmente conjuntos de dados moleculares em grande escala gerados por sequenciamento do genoma e outros dados de alto rendimento tecnologias experimentais (https://www.genome.jp/kegg/docs/release.html).

Outra forma de se trabalhar com dados obtidos de ensaios biológicos é agrupá-los em *clusters*. Clusterização(*clustering*) não é um tópico novo para estudos na área de Biologia Computacional. Na análise dos dados de expressão gênica, genes obtidos a partir de dados de *microarray*, formam *clusters* e os genes no mesmo agrupamento são considerados como ativadores da mesma função.

Vários algoritmos avançados de clusterização têm sido propostos (BOUTROS; OKEY, 2005), incluindo clusterização hierárquica (LAFOND-LAPALME et al., 2017) e *clustering by ensemble* (YU et al., 2014). Para evitar o ruído em dados de *microarray*, *bi-clustering* (ZHANG et al., 2017) foi empregado, sendo a seleção de características e seleção de amostras são realizadas ao mesmo tempo. Dados de sequenciamento de célula única funcionam de forma semelhante a dados de expressão gênica. Diferentes expressões gênicas foram agrupadas de acordo com diferentes tipos de células (AIBAR et al., 2017) (LI et al., 2017). No entanto, estes métodos focam em valores de expressão gênica ao invés de sequências de gene.

4.3 Levantamentos de dados

Os dados analisados são provenientes de uma análise de *microarray*, na qual foi medida a expressão de um total de 20.183 genes, expressos em células cancerígenas de mama, antes e depois do tratamento com um composto promissor para tratamento desse câncer (MIH2.4) (COSTA et al., 2020).

Neste experimento, 732 genes diferencialmente expressos foram identificados de um total de 20.183 genes com expressão medida. Estes foram identificados usando limites definidos pelo usuário. Neste experimento, o usuário escolheu um limite de 0,05 para significância estatística (p-valor) e uma mudança logarítmica de expressão com valor absoluto de pelo menos 2. Essas informações estão armazenadas em uma planilha de Excel.

4.4 Análise preliminar dos dados

Para a análise dos dados, efetuou-se um levantamento bibliográfico preliminar, bem como, estudos que permitam o processamento de dados. Dentre as ferramentas comumente utilizadas, podemos citar:

- Gene Ontology (GO) (http://geneontology.org/) e
- Kegg database (https://www.genome.jp/kegg/tool/map_pathway2.html).

Estes servem para análise dos genes diferencialmente expressos, envolvidos em processo biológico, componente celular e função molecular, utilizando o banco de dados DAVID, além da ferramenta *cluster* (YUAN et al., 2018). As metodologias aqui indicadas constituem um tratamento inicial para os problemas pesquisados, advindas de estudos prévios realizados. Novas abordagens podem ser incorporadas ou desenvolvidas posteriormente.

A primeira abordagem estudada trata-se da análise de agrupamento, ou *clustering*, nome dado para o grupo de técnicas computacionais cujo propósito consiste em separar objetos em grupos, baseando-se nas características que estes objetos possuem. A ideia básica consiste em colocar em um mesmo grupo objetos que sejam similares de acordo com algum critério pré-determinado.

O critério baseia-se normalmente em uma função de dissimilaridade, função esta que recebe dois objetos e retorna a distância entre eles. Os grupos determinados por uma métrica de qualidade devem apresentar alta homogeneidade interna e alta separação (heterogeneidade externa). Isto quer dizer que os elementos de um determinado conjunto devem ser mutuamente similares e, preferencialmente, muito diferentes dos elementos de outros conjuntos.

A análise de agrupamento é uma ferramenta útil para a análise de dados em muitas situações diferentes. Esta técnica pode ser usada para reduzir a dimensão de um conjunto de dados, reduzindo uma ampla gama de objetos à informação do centro do seu conjunto. Tendo em vista que a clusterização é uma técnica de aprendizado não supervisionado (quando o aprendizado é supervisionado, o processo é denominado de classificação), pode servir também para extrair características dos dados e desenvolver as hipóteses a respeito de sua natureza.

4.5 Processamento dos dados

Para o processamento dos dados brutos, obtidos de uma planilha no formato xls, com mais de 3216 genes resultantes de uma análise de *microarray*, foi utilizado o programa R versão 4.3.0 que pode ser baixado em g e instalado nos três sistemas operacionais principais *Windows, Mac, Unix/Linux* (Figura 5).

4.6 Lista de Pacotes de R complementares para a análise.

Figura 5 – Imagem inicial programa R.

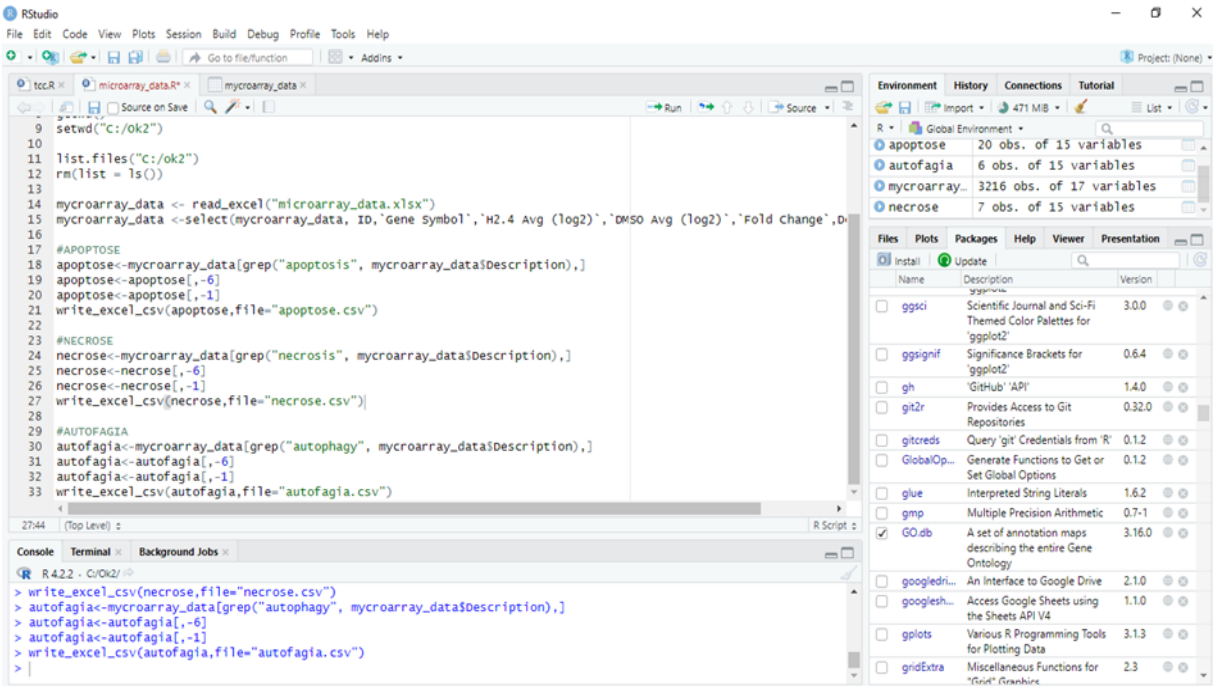


Tabela 1 – Lista de Pacotes de R complementares para a análise.

PACOTE	DESCRIÇÃO
affy	Ferramenta de análise de dados de expressão gênica em larga escala.
affydata	Conjunto de dados de *microarray* pré-processados e normalizados, que pode ser usado para testar e demonstrar o uso do pacote R affy.
affyPLM	Ferramenta para modelagem e análise de dados de *microarray* usando modelos lineares.
annaffy	Ferramenta para a anotação e visualização de dados de *microarray*.
annotate	Ferramenta para a anotação de genes e sequências biológicas. Fornece funções para a recuperação de informações de anotação de bancos de dados públicos, como GenBank e UniProt, e para a associação dessas informações com sequências biológicas.
Biobase	Ferramenta para a manipulação e análise de dados de biologia molecular, incluindo dados de *microarray* e sequenciamento de nova geração (NGS).

PACOTE	DESCRIÇÃO
Biostrings	Ferramenta para manipulação e análise de sequências biológicas, incluindo DNA, RNA e proteínas.
DESeq2	Ferramenta para a análise de expressão gênica diferencial em dados de RNA-Seq.
dplyr	Permite que o usuário execute operações comuns em bancos de dados, como filtragem, ordenação, agrupamento, seleção de colunas, adição de colunas calculadas, entre outras.
DynDoc	Ferramenta para a criação e visualização interativa de documentos dinâmicos. Ele fornece funções para a criação de documentos RMarkdown interativos que permitem a exploração de dados e a interação do usuário com gráficos e tabelas.
gcrma	Ferramenta para a normalização de dados de *microarray* utilizando o método de ajuste de modelo de sequência GCRMA (GeneChip Robust Multi-array Average).
genefilter	Ferramenta para filtragem de genes em análise de dados de expressão gênica.
geneplotter	Ferramenta para visualização de dados de expressão gênica. Fornece funções para a criação de gráficos de expressão gênica e para a visualização de dados de *microarray* e RNA-Seq.
GenomeInfoDb	Ferramenta para armazenamento e gerenciamento de informações sobre genomas e sequências genômicas.
GEOquery	Ferramenta para acesso e análise de dados do Gene Expression Omnibus (GEO), um dos maiores repositórios públicos de dados de expressão gênica.
ggpltot2	Ferramenta utilizada para criação de gráficos estatísticos com base em camadas. Possibilita a criação de gráficos de dispersão, histogramas, gráficos de barras, gráficos de linhas e mapas.
GO.db	Biblioteca R que contém informações sobre os termos e relações do Gene Ontology (GO).
hgu133plus2	Biblioteca R que contém informações sobre os probesets e genes associados ao chip de *microarray Affymetrix HG-U133 Plus* 2.0.

PACOTE	DESCRIÇÃO
hgu133plus2cdf	Biblioteca R que contém definições de arquivo de matriz de expressão (CDF) para o chip de *microarray Affymetrix* HG-U133 Plus 2.0.
hgu133plus2probe	Biblioteca hgu133plus2probe é uma biblioteca R que contém informações sobre os probes individuais usados no chip de *microarray Affymetrix* HG-U133 Plus 2.0.
IRanges	Fornece uma estrutura de dados para armazenar e manipular intervalos e conjuntos de intervalos.
lattice	Fornece uma estrutura para a criação de gráficos estatísticos de alta qualidade, com suporte para múltiplas variáveis.
limma	Fornece um conjunto de ferramentas para a análise de *microarray* e dados de sequenciamento de RNA.
marray	Fornece ferramentas para análise de *microarrays* de duas cores.
MASS	Fornece uma coleção de funções e conjuntos de dados para a análise estatística multivariada.
multtest	Fornece ferramentas para a correção de múltiplos testes estatísticos.
plier	Fornece ferramentas para pré-processamento e análise de dados de *microarray* de expressão gênica.
preprocessCore	Fornece ferramentas para pré-processamento de dados de *microarray* de expressão gênica.
readr	Fornece ferramentas para importar e exportar dados em diversos formatos de arquivos, incluindo CSV, TSV e arquivos de valores separados por espaço.
readxl	Ferramenta utilizada para importar dados de planilhas do *Microsoft* Excel.
Rmpfr	Fornece ferramentas para cálculos numéricos de alta precisão.
ROC	Fornece ferramentas para criação e análise de curvas ROC (Receiver Operating Characteristic).
S4Vectors	Fornece estruturas de dados e funções para manipulação eficiente de dados de biologia molecular em grande escala.
stringi	Ferramenta do R utilizada para manipulação de strings em vários formatos e encodings.

PACOTE	DESCRIÇÃO
stringr	Ferramenta do R utilizada para manipulação de strings, incluindo a busca, extração, substituição, concatenação e formatação de strings.
tibble	Oferece uma maneira mais moderna e eficiente de armazenar dados em formato tabular.
tidyverse	Consolida uma série de ferramentas que fazem parte do ciclo da ciência de dados ggplot2, dplyr, tidyr, purrr e readr.
vsn	Ferramenta de normalização de dados que é frequentemente usada em análise de microarray.
xfun	Coleção de funções auxiliares úteis para diversos tipos de projetos em R.
XML	Fornece funções para manipulação de dados em formato XML (Extensible Markup Language) em R.
xtable	Cria tabelas formatadas em LaTeX, HTML, Texto ou outros formatos.

Leitura da planilha e armazenamento na variável mycroarray_data.

>mycroarray_data <- read_excel(”microarray_data.xlsx”)

Seleção das colunas trabalha na variável:

>mycroarray_data <-select(mycroarray_data, ID,‘Gene Symbol‘,‘H2.4 Avg (log2)‘, ‘DMSO Avg (log2)‘, ‘Fold Change‘,’Description’)

Separação dados dos genes responsáveis pela morte celular por apoptose:

>apoptose<-mycroarray_data[grep(”apoptosis”,mycroarray_data$Description),]

>apoptose<-apoptose[,-6]

>apoptose<-apoptose[,-1]

>write_excel_csv(apoptose,file=”apoptose.csv”)

Separação dados dos genes responsáveis pela morte celular por necrose:

>necrose<-mycroarray_data[grep(”necrosis”, mycroarray_data$Description),]

>necrose<-necrose[,-6]

>necrose<-necrose[,-1]

>write_excel_csv(necrose,file=”necrose.csv”)

Separação dados dos genes responsáveis pela morte celular por autofagia:

autofagia<-mycroarray_data[grep(”autophagy”, mycroarray_data$Description),]

>autofagia<-autofagia[,-6]

>autofagia<-autofagia[,-1]

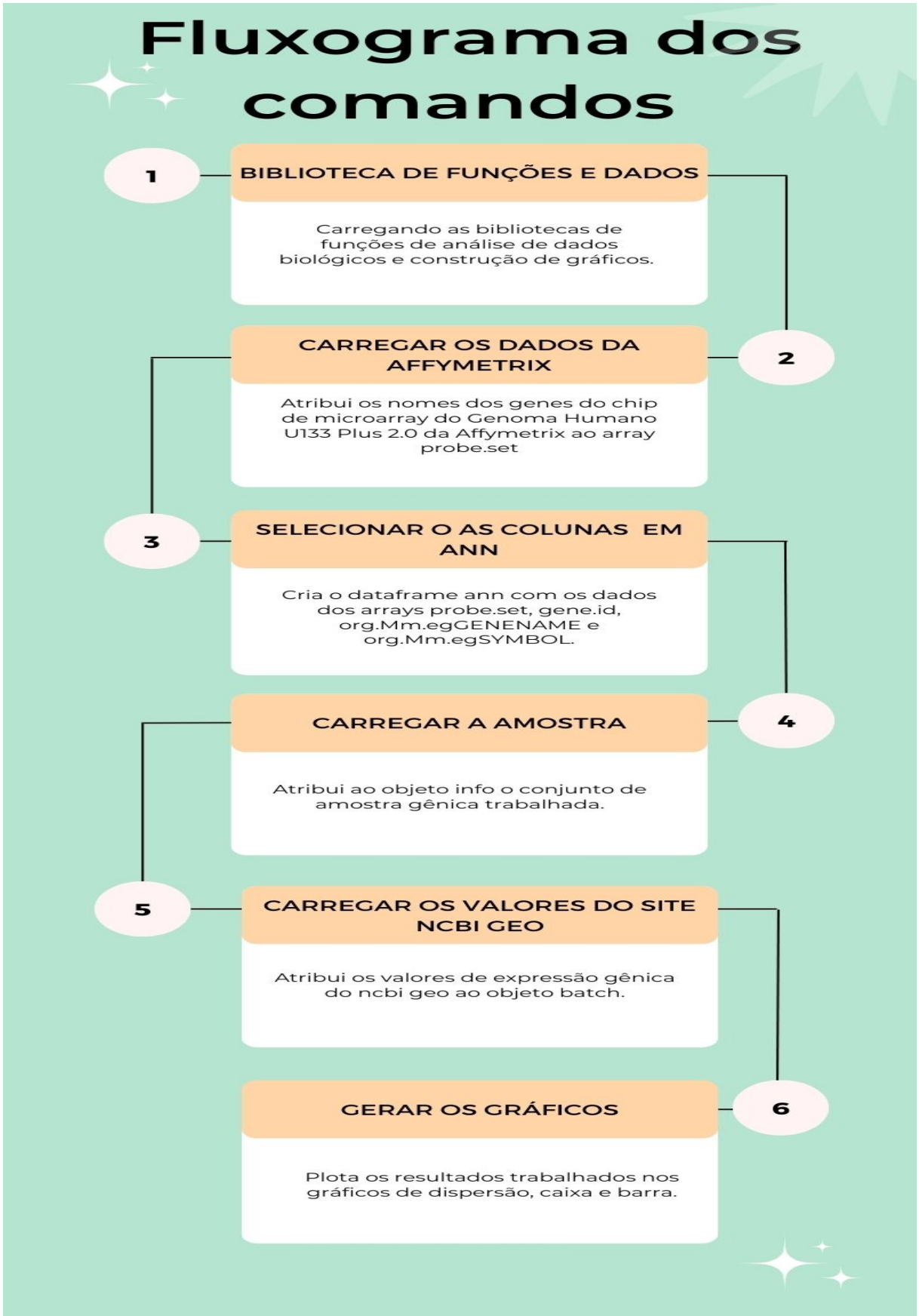
>write_excel_csv(autofagia,file=”autofagia.csv”)

O processamento, solicitado pela equipe de biólogos participantes do presente projeto e colaboradores da pesquisa relacionada à obtenção dos dados aqui utilizados, visam a análise dos dados e categorização dos genes envolvidos nos processos de morte celular (apoptose, necrose e autofagia), obtido no arquivo formato txt, para que pudessem ser utilizados no site: http://www1.heatmapper.ca/pairwise/.

4.7 *Script* da normalização de dados provenientes da análise de *Microarray*

Para o início do desenvolvimento e normalização dos dados, selecionamos no site http://bioconductor.org/packages/2.1/AffymetrixChip.html o conjunto de dados *Affymetrix Human Genome U133 Pus 2.0 Array Annotation data* (hgu133plus2), de acordo com o que foi descrito no trabalho de COSTA et al., 2020, trabalho que nos forneceu nossa fonte de estudo. Para a organização dos dados dos genes como níveis de expressão diferencialmente expressos em uma tabela de excel foi utilizado um conjunto de dados de células MCF-7 que estão depositadas no repositório público de dados genômicos funcionais NCBI (*National Center for Biotechnology Information*) *GEO acession viewer* presentes no site https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148537 referentes a linhagem MCF-7. Em seguida foi desenvolvido um *script* no programa R para facilitar a extração dos dados de *microarray* MCF-7, como também foi confeccionado um fluxograma Figura 6 para facilitar a compreensão do *script*.

Figura 6 – Fluxograma dos comandos.



Fonte: Produzido pelo próprio autor

4.7.1 Biblioteca de funções e dados

Instalar o Biocmanager:

>install.packages(”BiocManager”)

>library(BiocManager)

>BiocManager::install(version = ”3.17”)

Utilizar os comandos exemplo abaixo para instalar as bibliotecas da tabela 1.

>BiocManager::install(”affy”)

>library(affy)

4.7.2 Carregar os dados da Affymetrix

Extrai uma lista única de nomes de sondas (probes) de um conjunto de dados de *microarray*.

>probe.set<-unique(as.data.frame(hgu133plus2probe)$Probe.Set.Name)

Substitui as ocorrências finais de ”_st”nos nomes das sondas (probes) por ”_at”.

> probe.set [grep(”_st$”, probe.set)]<-paste(probe.set [grep(”_st$”,probe.set)],”at”, sep=”_”)

Cria um novo objeto chamado gene.id, que contém os nomes de genes correspondentes aos nomes de sondas no objeto.

> gene.id<-sub(”_at”,, probe.set)

4.7.3 Selecionar as colunas em ANN

> ann<-as.data.frame(matrix(nrow=length(gene.id), ncol=4))

Atribui nomes às dimensões de uma matriz de anotação (annotation matrix) chamada ann.

> dimnames(ann)<-list(probe.set, c(”ProbeSet”, ”GeneID”, ”Symbol”,”GeneName”))

Adiciona uma nova coluna na matriz de anotação ann contendo os nomes de sondas (probes) contidos no objeto ’probe.set’.

> ann$probeset<-probe.set

Adiciona uma nova coluna na matriz de anotação ann contendo os nomes de sondas (geneid) contidos no objeto ’geneid’.

> ann$GeneID<-gene.id

Tabela 2 – sampleInfo.txt

File name	Sample name	Group	BS
GSM2189.CEL	FVB_E12.5_1-2-3-m5	E12.5	1-2-3-m5
GSM2190.CEL	FVB_E12.5_4-5-6-m5	E12.5	4-5-6-m5
GSM2191.CEL	FVB_E12.5_7-8-9-m5	E12.5	7-8-9-m5
GSM2192.CEL	FVB_NN_1-2-m5	NN	1-2-m5
GSM2193.CEL	FVB_NN_7-8-m5	NN	7-8-m5
GSM2194.CEL	FVB_NN_9-10-m5	NN	9-10-m5
GSM2088.CEL	FVB_1w_801-m5	A1w	801-m5
GSM2178.CEL	FVB_1w_804-m5	A1w	804-m5
GSM2179.CEL	FVB_1w_805-m5	A1w	805-m5
GSM2183.CEL	FVB_4w_11293-m5	A4w	11293-m5
GSM2184.CEL	FVB_4w_11294-m5	A4w	11294-m5
GSM2185.CEL	FVB_4w_11295-m5	A4w	11295-m5
GSM2334.CEL	FVB_3m_1f-m5	A3m	1f-m5
GSM2335.CEL	FVB_3m_2f-m5	A3m	2f-m5
GSM2336.CEL	FVB_3m_3f-m5	A3m	3f-m5
GSM2186.CEL	FVB_5m_731m-m5	A5m	731m-m5
GSM2187.CEL	FVB_5m_732m-m5	A5m	732m-m5
GSM2188.CEL	FVB_5m_733m-m5	A5m	733m-m5
GSM2180.CEL	FVB_1y_511m-m5	A1y	511m-m5
GSM2181.CEL	FVB_1y_5m-m5	A1y	5m-m5
GSM2182.CEL	FVB_1y_6m-m5	A1y	6m-m5
GSM2337.CEL	FVB_1y_529f-m5	A1y	529f-m5
GSM2338.CEL	FVB_1y_530f-m5	A1y	530f-m5
GSM2339.CEL	FVB_1y_544f-m5	A1y	544f-m5

Adiciona uma nova coluna na matriz de anotação ’ann’ contendo os nomes dos genes correspondentes a cada sonda.

> ann$GeneName<-unlist(unlist(as.list(org.Mm.egGENENAME)))[gene.id]

Adiciona uma nova coluna na matriz de anotação ’ann’ contendo os simbolo correspondentes a cada sonda.

> ann$Symbol<-unlist(unlist(as.list(org.Mm.egSYMBOL)))[gene.id]

Cria o diretório ”obj”.

> if(!file.exists(”obj”)) dir.create(”obj”)

Salva um objeto ann no arquivo obj/ann.RData.

> save(ann, file=”obj/ann.RData”)

> tail(ann)

4.7.4 Carregar a amostra

Salva tabela em info

> info<-read.table(”sampleInfo.txt”,header = T)

Transforma a variável Group do objeto info em um fator, com níveis especificados na ordem fornecida.

> info$Group<-factor(info$Group,levels=c(”E12.5”, ”NN”, ”A1w”, ”A4w”, ”A3m”, ”A5m”, ”A1y”))

4.7.5 Carregar os valores do site NCBI GEO

Obtêm o arquivo complementar ”GSE148537”

> getGEOSuppFiles(”GSE148537”)

Descompacta um arquivo tar em um diretório especificado.

> untar(”GSE148537/GSE148537_raw.TAR”,exdir=’data/’)

Carrega os arquivos .cel EM ”batch”.

> batch<-ReadAffy(celfile.path=”data/”)

Normaliza os dados de *microarray* e remover variações técnicas.

> normalized.data <- rma(batch)

Obtêm estimativas de expressão.

> normalized.expr <- as.data.frame(exprs(normalized.data))

Obtêm mapa de ids de probe para símbolos de gene.

> gse<-getGEO(”GSE148537”,GSEMatrix = TRUE)

buscar dados de recurso para obter mapeamento de id - símbolo de gene.

> feature.data<-gse$GSE148537_series_matrix.txt.gz@featureData@data

> feature.data<-feature.data[,c(1,11)]

Realizar um inner join entre a tabela de dados de expressão gênica normalizada.

> normalized.expr<- normalized.expr %>%

+ rownames_to_column(var = ’ID’) %>%

+ inner_join(., feature.data, by=’ID’)

> cdfName(batch)

> batch@cdfName<-”hgu133plus2”

> save(batch, file=”obj/batch.RData”)

Obtêm o pré-processamento de dados de *microarray* aos dados brutos contidos no objeto “batch”.

4.7.6 Gerar os gráficos

> eset<-justplier(batch, normalize=t)

Extrai a matriz de expressão gênica do objeto ’eset’

> exp<-exprs(eset)

Criar um gráfico de dispersão das combinações de todas as variáveis de um conjunto de dados.

> pairs(exp[,1:3])

Ajusta os dados de ’exp’.

exp<-log2(2êxp+16)

Cria um gráfico de dispersão das combinações dos valores de expressão dos três primeiros genes.

> pairs(exp [,1:3])

> save(exp, file=”obj/exp.rdata”)

Cria uma matriz contendo os valores de chamadas de presença (p), ausência (a) ou indefinido (m) para cada uma das sondas do *microarray*.

> calls<-exprs(mas5calls(batch))

Gera um gráfico de distribuição dos valores de expressão.

> boxplot(exp calls)

Soma as ocorrências ”p”em uma coluna.

>row.calls<-rowsums(calls==”P”)

Retorna a quantidade de ocorrências.

table_calls <- table(row.calls)

Plota a quantidade de ocorrencias em um gráfico de barras.

> barplot(table_calls)

Retorna a quantidade de ocorrências.

> table_calls <- table(row.calls)

4.8 Conclusão

O presente capítulo trouxe uma descrição da metodologia utilizada, a qual envolveu a compreensão do conjunto de dados, das consultas de interesse do grupo de biólogos que trabalharam no experimento, além da apresentação dos *scripts* em R desenvolvidos para tais questões. Os resultados obtidos a partir de tais estudos são apresentados e discutidos no capítulo a seguir.

5 Resultados e Discussão

A partir da definição dos dados biológicos, e para uma melhor compreensão do problema de pesquisa bem como de técnicas a serem utilizadas para abordá-lo, foram realizados estudos na área de biologia molecular, bioinformática, bem como, métodos computacionais de clusterização.

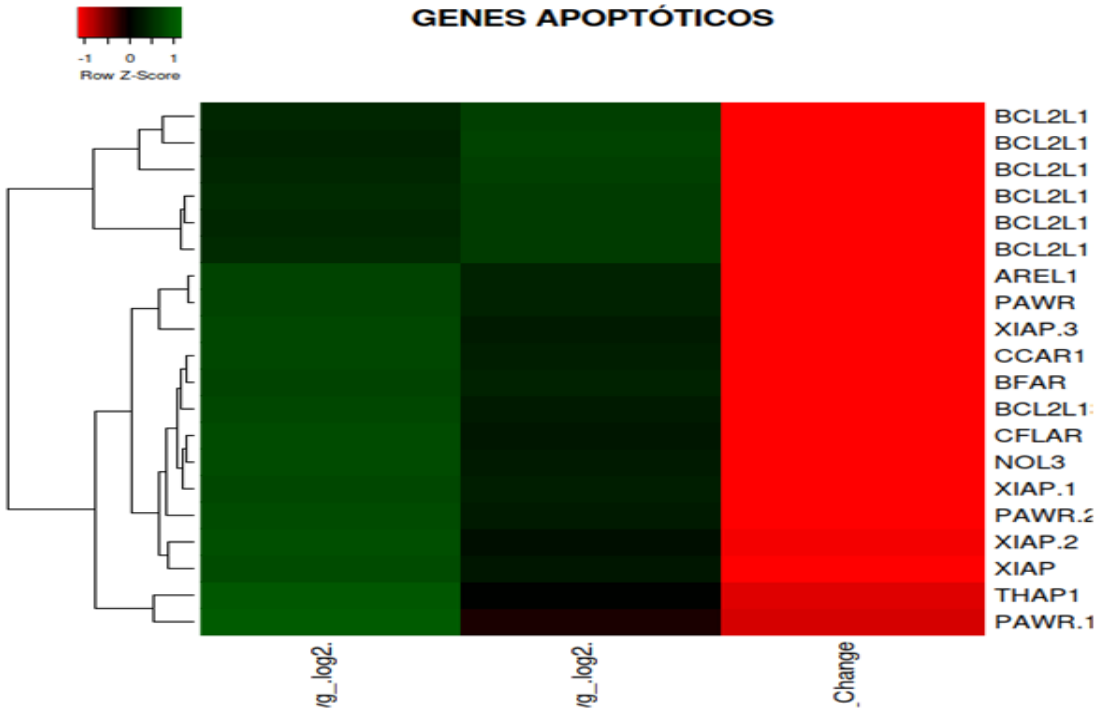
Como resultados, após análise de ferramentas advindas da biologia computacional que possam auxiliar na interpretação de dados biológicos foram definidos alguns programas que poderão ser usados na análise dos dados, são eles: para o alinhamento de sequências pode ser utilizado o *Blast* e *Clustal W*, modelagem de proteínas por homologia usando Modeller e biblioteca *Biopython*, entendimento de técnicas de clusterização *K-means*, clusterização hierárquica e análise de dados no programa R.

Para possibilitar, assim, a categorização dos genes envolvidos no processo de morte apopitoticas, necrótica e autofagica, processos que foram solicitados pelo grupo de pesquisa que nos cedeu os dados. Como a análise de *microarray* gerou como resultado uma tabela do tipo xls, com um grande volume dados, foi utilizado o programa R para que fossem selecionados nessa tabela, os genes envolvidos em processo de morte necrótica, apoptótica e autofágica. Após seleção dos dados, observou-se que dos 732 genes diferencialmente expressos, vinte genes estão envolvidos na cascata de morte apoptótica, seis envolvidos na morte autofágica e sete envolvidos na morte necrótica.

Após a seleção dos grupos de genes envolvidos no processo de morte celular, apoptotica, necrótica e autofágica, foram construídos heatmaps como forma de visualização dos dados revelando estruturas de agrupamento hierárquico presentes nas colunas e nas linhas de uma matriz. Essa matriz é representada por uma cor pertencentes a uma escala de cores pré-estabelecida (FERRETTI, 2015) A clusterização do tipo *Heatmaps* tornou-se uma ferramenta essencial para gerar novas hipóteses a partir de dados de expressão gênica, além de ser considerado um dos primeiros passos na análise da expressão gênica (D'HAESELEER, 2005). Os nomes dos genes são exibidos o lado direito da figura e os valores dos níveis de expressão estão demonstrados na cor coluna vermelha e em valores de *fold-change* (diferença da expressão gênica para os valores antes e depois do tratamento das células cancerígenas com o composto mesoiônico, Figuras 7, 8 e 9.

A bioinformática e a biologia computacional consistem de um complexo e multidisciplinar vasto, onde inúmeras ferramentas ao longo dos anos foram desenvolvidas para analisar quantidades crescentes de dados (GOUJON et al., 2010). Experimentos

Figura 7 – *Heatmaps* da expressão gênica de genes que estão envolvidos no processo de morte apoptótico.

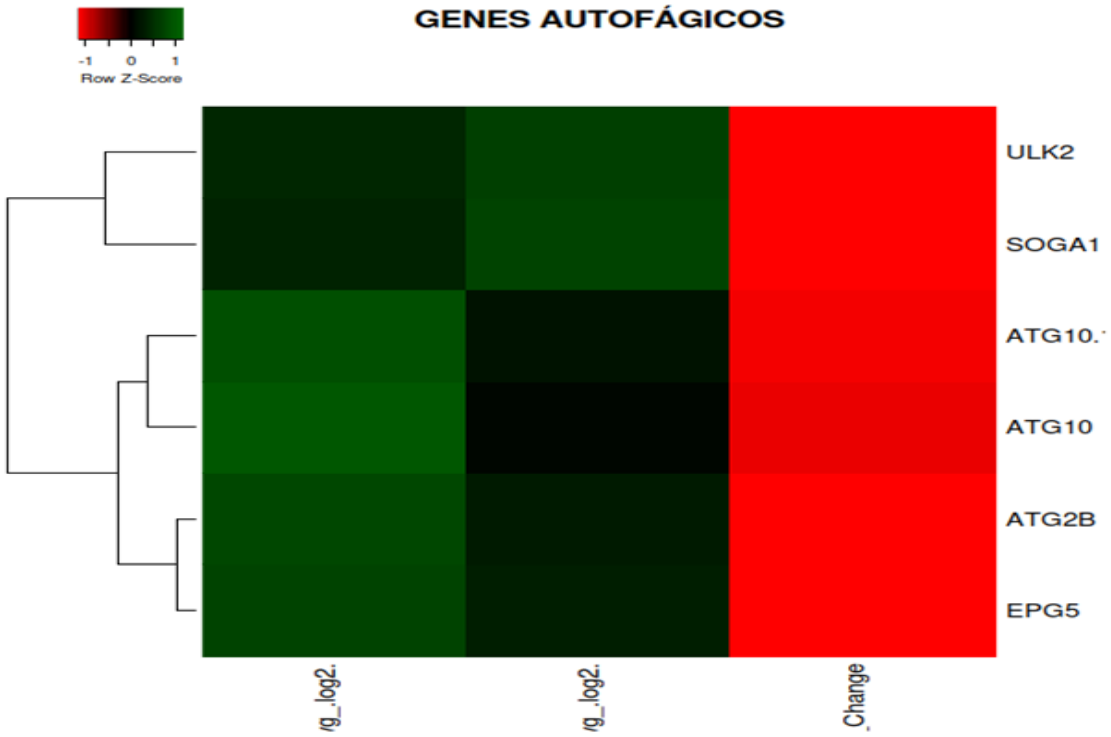


Fonte: Produzido pelo próprio autor

genômicos de alto rendimento, que envolvem o uso de *microarrays* ou tecnologias de sequenciamento de última geração, são usados em diversas áreas da biologia molecular para descobrir genes envolvidos em processos biológicos particulares. Para melhor compreensão dos resultados, as análises de perspectiva global e ferramentas de visualização intuitivas são necessárias para extrair informações relevantes (PEREZ-DIEZ; MORGUN; SHULZHENKO, 2007). Uma forma de demonstrar estes dados, são os *Heatmaps* que por representações gráficas de dados, onde os valores em uma matriz são representados através de cores (PEREZ-DIEZ; MORGUN; SHULZHENKO, 2007). Uma série de ações podem ser realizadas por avaliação dos *Heatmaps*, explorar e interpretar os resultados de forma eficaz, como pesquisar, filtrar por valor ou rótulo, agrupar o *Heatmap*, classificar linhas e colunas por critérios diferentes e etc, facilitando assim o trabalhos dos pesquisadores na interpretação dos dados gerados.

Após a obtenção dos dados selecionados da tabela xls e construçãos dos *clusteres*, foi utilizado o banco de dados http://bioconductor.org/packages/2.1/AffymetrixChip.html para a seleção do conjunto de dados *Affymetrix Human Genome U133 Pus 2.0 Array* Annotation data (hgu133plus2) para entendermos o processo de normalização dos dados brutos provenientes de um análise *microarrays* similar ao da nossa fonte de estudo. Diante disso foi construido gráficos de dispersão par a par de três matrizes, gráfico de dispersão par após correção logarítmica, gráficos de dispersão com valores

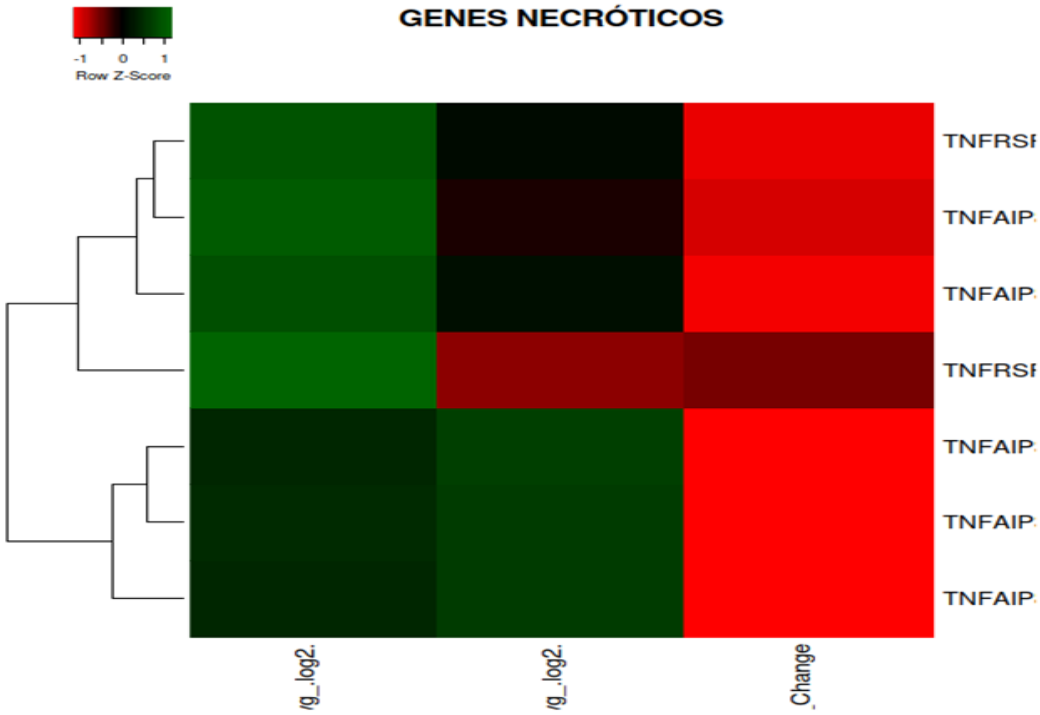
Figura 8 – *Heatmaps* da expressão gênica genes que estão envolvidos no processo de morte autofágico.



Fonte: Produzido pelo próprio autor

de presença (P), ausência (A) ou indefinido (M) e gráficos de variância que servem para avaliar estatisticamente o grande volume de dados brutos, obtidos da análise. De acordo com Mazza, ao analisar um gráfico de dispersão é importante observar a densidade de pontos em cada painel e identificar possíveis padrões visuais (MAZZA, 2009). No caso do exemplo que foi gerado na Figura 10 , observamos que não há uma distribuição uniforme de pontos em todos os painéis, sem uma tendência clara de aumento ou diminuição. Isso sugere que não há uma forte relação linear entre as três matrizes, o que é consistente com o fato de que as matrizes têm uma relação fraca. No entanto, é possível que haja outras formas de relação não linear entre as matrizes que não são capturadas por um gráfico de dispersão par a par. Por exemplo, pode haver uma relação quadrática entre duas matrizes, que não seria identificada por um gráfico de dispersão simples. Já ao analisarmos o gráfico Figura 11, podemos observar que há uma relação forte entre X e Y, com pontos agrupados em uma diagonal ascendente. Isso sugere que há uma relação linear positiva entre as matrizes X e Y. Além disso, a correção logarítmica torna a densidade de pontos mais clara, permitindo uma melhor visualização da distribuição dos pontos. No entanto, é importante lembrar que o uso desta correção pode afetar a interpretação dos dados, uma vez que a adição de valores aleatórios pode alterar a posição relativa dos pontos (SHISHKIN; SOKER, 2021). É importante avaliar se os valores para a correção é apropriado para o conjunto de dados

Figura 9 – *Heatmaps* da expressão de genes envolvidos no processo de morte necrótica.

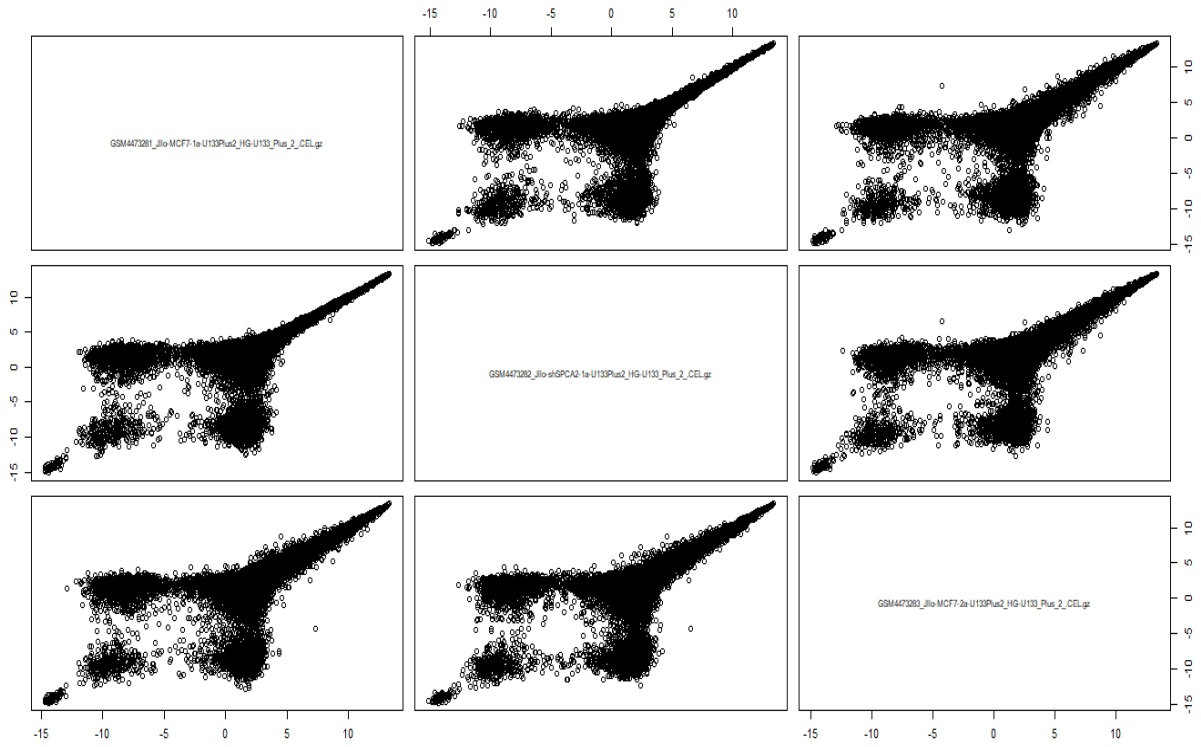


Fonte: Produzido pelo próprio autor

e o objetivo da análise (SHISHKIN; SOKER, 2021). Além disso, é sempre importante realizar uma análise mais profunda dos dados para confirmar qualquer relação aparente identificada no gráfico de dispersão.

Quando os gráficos de distribuição de valores de chamadas de presença, conforme apresentado na Figura 12, ausência ou indefinido para cada uma das sondas do *microarray* apresentam chamadas correlacionadas com o nível de expressão, mas sem uma diferença clara entre os três grupos, isso sugere que a expressão gênica pode ser uniforme em todas as amostras. Essa uniformidade pode ser causada por vários fatores, como condições experimentais semelhantes, baixa variabilidade biológica ou baixa sensibilidade do método de detecção. Quando a variação é baixa, é comum ver valores de sonda consistentes em todas as amostras, resultando em uma sobreposição dos três grupos P, A e M no gráfico de distribuição. No entanto, é importante lembrar que a ausência de uma diferença clara entre os grupos de chamadas não significa necessariamente que não exista diferença de expressão gênica entre as amostras. Pode haver diferenças sutis na expressão que não são detectadas pelo método de *microarray*. Nesses casos, é importante realizar uma análise mais aprofundada dos dados, incluindo a normalização dos dados, análise de expressão diferencial e validação experimental, como a análise de RT-PCR(PCR em tempo real). Essas técnicas podem ajudar a identificar diferenças significativas na expressão gênica, mesmo quando

Figura 10 – Gráfico de dispersão.



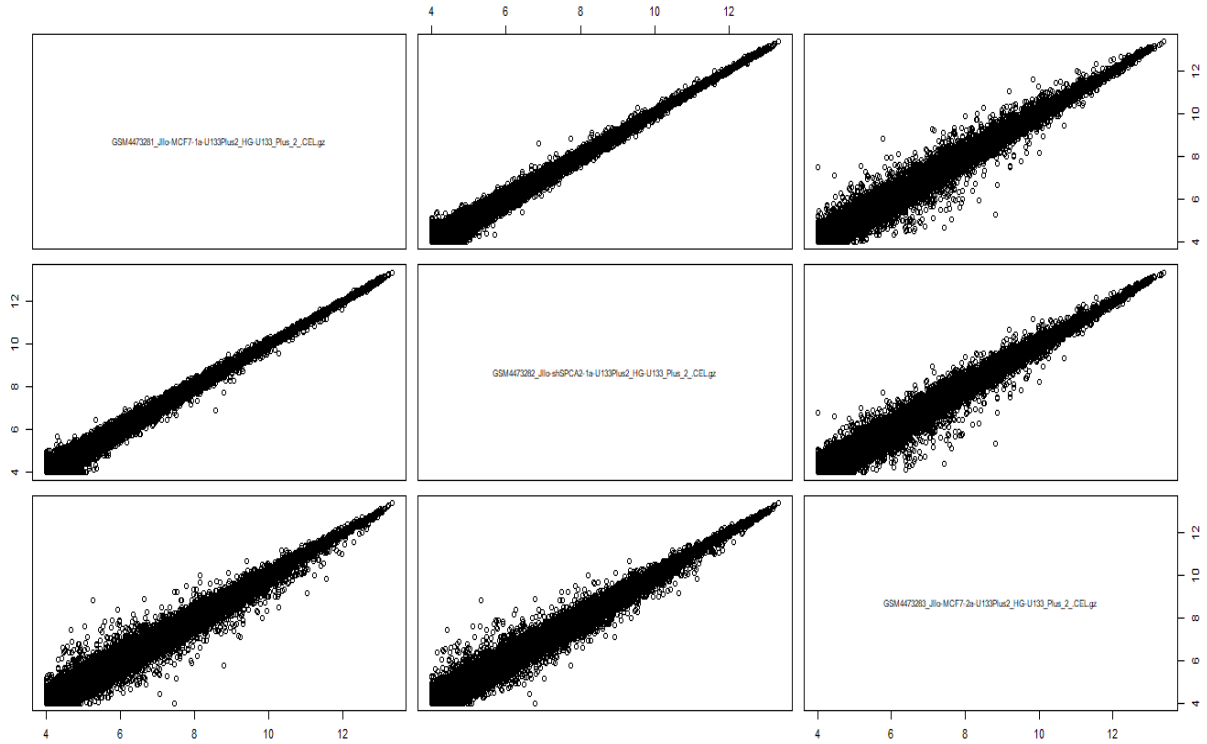
Fonte: Produzido pelo próprio autor

o gráfico de distribuição de chamadas não é conclusivo. A técnica de RT-PCR, foi desenvolvida baseiando-se em estudos envolvendo expressão gênica e trata-se de um método de análise para medir o acúmulo dos produtos de PCR a medida que vão sendo produzidos, e dessa maneira,a quantificar as cópias de mRNA dos genes em estudo. Essa técnica é bastante sensivel e tem grande confiabilidade, sendo um excelente instrumento de quantificação de expressão gênica, Além disso atualmente é considerado o método mais apropriado para confirmar os dados gerados por *microarray* (PROVENZANO; MOCELLIN, 2007).

O gráfico de variância em barras Figura 13 é uma ferramenta valiosa para comparar a variação entre diferentes grupos de dados. Ele pode ajudar a identificar padrões e tendências nos dados e fornecer informações úteis para análises estatísticas mais avançadas.

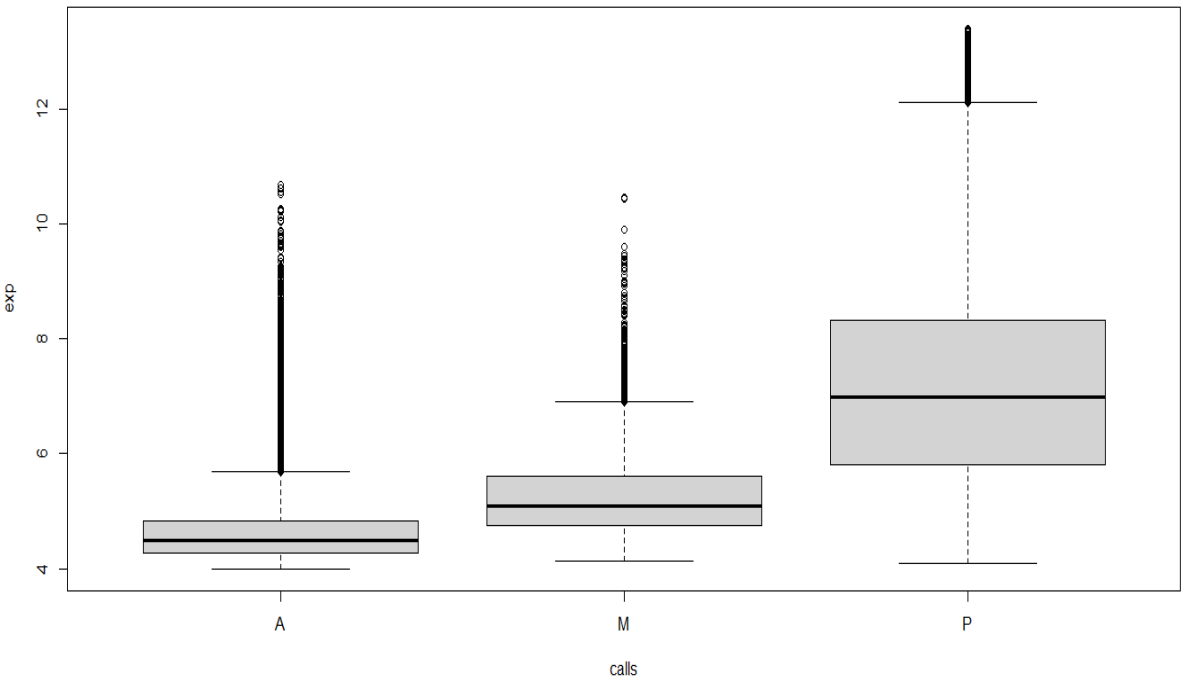
Os resultados advindos do presente trabalho se mostraram promissores e infundiu trabalhos futuros. As conclusões da pesquisa descrita são apresentadas no capítulo a seguir.

Figura 11 – Gráfico de dispersão após correção.



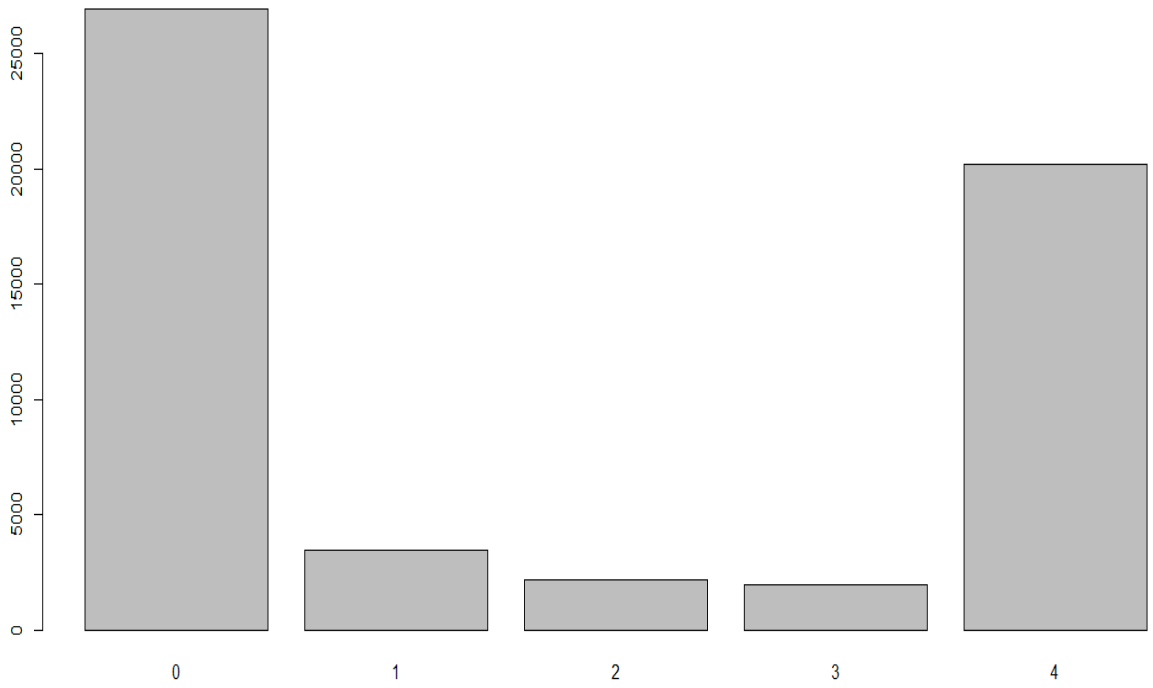
Fonte: Produzido pelo próprio autor

Figura 12 – Gráfico de dispersão com valores de presença(P), ausência(A) ou indefinido(M).



Fonte: Produzido pelo próprio autor

Figura 13 – Gráfico de barras.



Fonte: Produzido pelo próprio autor

6 Conclusões e Trabalhos Futuros

A Ciência da Computação estabeleceu-se como um meio fundamental para pesquisas na área de Biologia Molecular. A utilização de *microarrays* tem se expandido devido às diversas aplicações em biologia e medicina, propiciando gerar uma quantidade significativa de dados utilizados em análises genômicas.

O estudo realizado neste trabalho, indicou a complexidade computacional envolvida na análise de *microarrays*, quer seja no gerenciamento, investigação ou mesmo interpretação dos resultados, para não especialistas em computação. Observou-se que forma de análise depende dos dados envolvidos no experimento, portanto, a definição do conjunto de dados foi efetuada no início da pesquisa. Assim, o estudo utilizou dados inéditos de uma pesquisa recente sobre desenvolvimento de fármaco para câncer de mama. Considerando as especificidades da análise, foi desenvolvida uma ferramenta que facilitou a busca por termos fundamentais na análise dos resultados do experimento, fornecendo, além do resultado numérico, uma representação gráfica, denominada *heatmap*, gerada a partir de técnicas de clusterização, facilitando a interpretação das informações obtidas.

Diante do que foi exposto observamos que os métodos computacionais podem contribuir para a evolução e aprimoramento de outras ciências, como as ciências biológicas. O problema abordado nesse trabalho, corroborou para detecção e seleção de genes alvo, os quais serão analisados em uma fase posterior da pesquisa. Uma outra contribuição que está em desenvolvimento, visa acrescentar outros tipos de análises de *microarray*, buscando facilitar a análise de resultados deste tipo de experimento por pesquisadores não especialistas em computação.

Referências

AGAPITO, G.; ARBITRIO, M. Microarray data analysis protocol. In: *Methods in Molecular Biology*. New York, NY: Springer US, 2022, (Methods in molecular biology (Clifton, N.J.)). p. 263–271. Citado 3 vezes nas páginas 13, 14 e 23.

AIBAR, S. et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, Nature Publishing Group US New York, v. 14, n. 11, p. 1083–1086, 2017. Citado na página 24.

ALVES, S. M. A bioinformática e sua importância para a biologia molecular. *Revista Brasileira de Educação e Saúde*, v. 3, n. 4, p. 18–25, 2013. Citado na página 15.

ARAÚJO, N. D. de et al. A era da bioinformática: seu potencial e suas implicações para as ciências da saúde. *Estudos de biologia*, v. 30, n. 70/72, 2008. Citado 3 vezes nas páginas 12, 15 e 17.

AZAD, R. K.; SHULAEV, V. Metabolomics technology and bioinformatics for precision medicine. *Brief. Bioinform.*, Oxford University Press (OUP), v. 20, n. 6, p. 1957–1971, nov. 2019. Citado na página 12.

BALL, G. H.; HALL, D. J. A clustering technique for summarizing multivariate data. *Behavioral science*, Wiley Online Library, v. 12, n. 2, p. 153–155, 1967. Citado na página 19.

BAO, R. et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer informatics*, SAGE Publications Sage UK: London, England, v. 13, p. CIN–S13779, 2014. Citado 2 vezes nas páginas 13 e 16.

BIRNEY, E. The international human genome project. *Hum. Mol. Genet.*, Oxford University Press (OUP), v. 30, n. R2, p. R161–R163, out. 2021. Citado na página 12.

BOUTROS, P. C.; OKEY, A. B. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. *Briefings in bioinformatics*, Henry Stewart Publications, v. 6, n. 4, p. 331–343, 2005. Citado na página 24.

COLOMBO, J.; RAHAL, P. A tecnologia de microarray no estudo do câncer de cabeça e pescoço. *Revista Brasileira de Biociências*, v. 8, n. 1, 2010. Citado na página 23.

COSTA, L. A. de M. et al. Mechanistic studies of cytotoxic activity of the mesoionic compound mih 2.4 bl in mcf-7 breast cancer cells. *Oncology letters*, Spandidos Publications, v. 20, n. 3, p. 2291–2301, 2020. Citado 3 vezes nas páginas 13, 14 e 24.

D'HAESELEER, P. How does gene expression clustering work? *Nature biotechnology*, Nature Publishing Group US New York, v. 23, n. 12, p. 1499–1501, 2005. Citado na página 37.

ESTEVES, G. H. *Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado*. Tese (Doutorado) — Universidade de São Paulo, 2007. Citado na página 17.

FARIAS, A.; CHACON, P.; SILVA, N. da. A bioinformática como ferramenta de formação de recursos humanos no ifrn. *HOLOS*, Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, v. 6, p. 113–123, 2012. Citado na página 15.

FERRETTI, Y. *Ferramenta computacional para análise integrada de dados clínicos e biomoleculares*. Tese (Doutorado) — Universidade de São Paulo, 2015. Citado na página 37.

FERTIG, E. J.; SLEBOS, R.; CHUNG, C. H. Application of genomic and proteomic technologies in biomarker discovery. *American Society of Clinical Oncology Educational Book*, American Society of Clinical Oncology, v. 32, n. 1, p. 377–382, 2012. Citado 2 vezes nas páginas 13 e 16.

GARRIDO-CARDENAS, J. et al. DNA sequencing sensors: An overview. *Sensors (Basel)*, MDPI AG, v. 17, n. 3, p. 588, mar. 2017. Citado na página 13.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, BioMed Central, v. 5, n. 10, p. 1–16, 2004. Citado na página 20.

GIACHETTO, P. A tecnologia de microarranjos na identificação de genes de interesse na bovinocultura. Campinas: Embrapa Informática Agropecuária, 2010., 2010. Citado 3 vezes nas páginas 17, 19 e 20.

GOLLUB, J.; SHERLOCK, G. [10] clustering microarray data. *Methods in enzymology*, Elsevier, v. 411, p. 194–213, 2006. Citado 2 vezes nas páginas 18 e 20.

GONÇALVES, M. L. et al. Classificação não-supervisionada de imagens de sensores remotos utilizando redes neurais auto-organizáveis e métodos de agrupamentos hierárquicos. *Revista Brasileira de Cartografia*, v. 60, n. 1, p. 17–29, 2008. Citado na página 19.

GOUJON, M. et al. A new bioinformatics analysis tools framework at embl–ebi. *Nucleic acids research*, Oxford University Press, v. 38, n. suppl_2, p. W695–W699, 2010. Citado na página 37.

GRIFFITHS, A. J. et al. Introdução à genética. In: *Introdução à genética*. [S.l.: s.n.], 2006. p. 743–743. Citado na página 12.

HOOD, L. Systems biology: integrating technology, biology, and computation. *Mechanisms of ageing and development*, Elsevier, v. 124, n. 1, p. 9–16, 2003. Citado 2 vezes nas páginas 12 e 15.

JOSHI, A. et al. Systems biology in cardiovascular disease: a multiomics approach. *Nat. Rev. Cardiol.*, Springer Science and Business Media LLC, v. 18, n. 5, p. 313–330, maio 2021. Citado na página 12.

KELL, D. B. The virtual human: towards a global systems biology of multiscale, distributed biochemical network models. *IUBMB life*, Wiley Online Library, v. 59, n. 11, p. 689–695, 2007. Citado na página 16.

LAFOND-LAPALME, J. et al. A new method for decontamination of de novo transcriptomes using a hierarchical clustering algorithm. *Bioinformatics*, Oxford University Press, v. 33, n. 9, p. 1293–1300, 2017. Citado na página 24.

LESK, A. M. *Introdução à bioinformática*. [S.l.]: Artmed, 2008. Citado na página 17.

LI, X. et al. Network embedding-based representation learning for single cell rna-seq data. *Nucleic acids research*, 2017. Citado na página 24.

MAZZA, R. *Introduction to information visualization*. [S.l.]: Springer Science & Business Media, 2009. Citado na página 39.

NIEMEYER, C. M.; BLOHM, D. Dna microarrays. *Angewandte Chemie International Edition*, Wiley Online Library, v. 38, n. 19, p. 2865–2869, 1999. Citado na página 22.

PELLENZ, F. M.; CRISPIM, D.; ASSMANN, T. S. Systems biology approach identifies key genes and related pathways in childhood obesity. *Gene*, Elsevier, v. 830, p. 146512, 2022. Citado na página 12.

PEREZ-DIEZ, A.; MORGUN, A.; SHULZHENKO, N. Microarrays for cancer diagnosis and classification. *Microarray technology and cancer gene profiling*, Springer, p. 74–85, 2007. Citado 2 vezes nas páginas 23 e 38.

PERVEZ, M. T. et al. A comprehensive review of performance of next-generation sequencing platforms. *Biomed Res. Int.*, Hindawi Limited, v. 2022, p. 3457806, set. 2022. Citado na página 13.

PRIYAMVADA, P. et al. A comprehensive review on genomics, systems biology and structural biology approaches for combating antimicrobial resistance in ESKAPE pathogens: computational tools and recent advancements. *World J. Microbiol. Biotechnol.*, Springer Science and Business Media LLC, v. 38, n. 9, p. 153, jul. 2022. Citado na página 12.

PROVENZANO, M.; MOCELLIN, S. Complementary techniques: validation of gene expression data by quantitative real time pcr. *Microarray Technology and Cancer Gene Profiling*, Springer, p. 66–73, 2007. Citado na página 41.

QUACKENBUSH, J. Computational analysis of microarray data. *Nature reviews genetics*, Nature Publishing Group UK London, v. 2, n. 6, p. 418–427, 2001. Citado na página 17.

REIMERS, M.; CAREY, V. J. [8] bioconductor: an open source framework for bioinformatics and computational biology. *Methods in enzymology*, Elsevier, v. 411, p. 119–134, 2006. Citado na página 20.

ROEHRS, M. et al. Retinol e carotenóides em pacientes hemodialisados e seus reflexos fisiopatológicos. Universidade Federal de Santa Maria, 2009. Citado na página 16.

RUSSO, G.; ZEGAR, C.; GIORDANO, A. Advantages and limitations of microarray technology in human cancer. *Oncogene*, Nature Publishing Group, v. 22, n. 42, p. 6497–6507, 2003. Citado na página 22.

SHISHKIN, D.; SOKER, N. Supplying angular momentum to the jittering jets explosion mechanism using inner convection layers. *Monthly Notices of the Royal Astronomical Society: Letters*, Oxford University Press, v. 508, n. 1, p. L43–L47, 2021. Citado 2 vezes nas páginas 39 e 40.

SMEDLEY, D. et al. Biomart–biological queries made easy. *BMC genomics*, BioMed Central, v. 10, n. 1, p. 1–12, 2009. Citado na página 23.

SUBRAMANIAN, I. et al. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, SAGE Publications Sage UK: London, England, v. 14, p. 1177932219899051, 2020. Citado 2 vezes nas páginas 13 e 16.

VIEIRA, L. S.; LAUBENBACHER, R. C. Computational models in systems biology: standards, dissemination, and best practices. *Curr. Opin. Biotechnol.*, Elsevier BV, v. 75, n. 102702, p. 102702, jun. 2022. Citado na página 12.

VIJAYAKUMAR, S. et al. Optimization of multi-omic genome-scale models: Methodologies, hands-on tutorial, and perspectives. In: *Methods in Molecular Biology*. New York, NY: Springer New York, 2018, (Methods in molecular biology (Clifton, N.J.)). p. 389–408. Citado na página 13.

WESTERHOFF, H. V.; PALSSON, B. O. The evolution of molecular biology into systems biology. *Nature biotechnology*, Nature Publishing Group UK London, v. 22, n. 10, p. 1249–1252, 2004. Citado na página 16.

YU, Z. et al. Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles. *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, v. 11, n. 4, p. 727–740, 2014. Citado na página 24.

YUAN, M. et al. Bioinformatics analysis of methylation in cervical adenocarcinoma in xinjiang, china. *Medicine*, Wolters Kluwer Health, v. 97, n. 35, 2018. Citado na página 25.

ZHANG, Y. et al. Application of computational biology and artificial intelligence in drug design. *Int. J. Mol. Sci.*, MDPI AG, v. 23, n. 21, p. 13568, nov. 2022. Citado na página 12.

ZHANG, Y. et al. Qubic: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics*, Oxford University Press, v. 33, n. 3, p. 450–452, 2017. Citado na página 24.