UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CARLOS RENAN MOREIRA

MINING_RNA: SISTEMA WEB PARA MINERAÇÃO DE DADOS EM ESTUDOS TRANSCRIPTÔMICOS A PARTIR DE MICROARRANJOS

Mossoró-RN

2021

CARLOS RENAN MOREIRA

MINING_RNA: SISTEMA WEB PARA MINERAÇÃO DE DADOS EM ESTUDOS TRANSCRIPTÔMICOS A PARTIR DE MICROARRANJOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, como requisito para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Profª Dra. CICÍLIA RAQUEL MAIA LEITE
Coorientador: Profª Dra. CHRISTINA PACHECO SANTOS MARTIN

Mossoró-RN

2021

Catalogação da Publicação na Fonte.
Universidade do Estado do Rio Grande do Norte.

M838m Moreira, Carlos Renan
MINING_RNA: Sistema web para mineração de dados em estudos transcriptômicos a partir de microarranjos. / Carlos Renan Moreira. - Mossoró, 2021.
56p.

Orientador(a): Profa. Dra. Cicília Raquel Maia Leite.
Coorientador(a): Profa. Dra. Christina Pacheco Santos Martin.
Dissertação (Mestrado em Programa de Pós-Graduação em Ciência da Computação). Universidade do Estado do Rio Grande do Norte.

1. Bioinformática. 2. Microarranjo. 3. GEO. 4. Sistema WEB. 5. Mineração de Dados. I. Maia Leite, Cicília Raquel. II. Universidade do Estado do Rio Grande do Norte. III. Título.

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pela Diretoria de Informatização (DINF), sob orientação dos bibliotecários do SIB-UERN, para ser adaptado às necessidades da comunidade acadêmica UERN.

CARLOS RENAN MOREIRA

MINING_RNA: SISTEMA WEB PARA MINERAÇÃO DE DADOS EM ESTUDOS TRANSCRIPTÔMICOS A PARTIR DE MICROARRANJOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 05 / 03 / 2021

Profa. Dra. Cicilia Raquel Maia Leite
Orientador e Presidente

Dra. Christina Pacheco Santos Martin
Coorientadora - Universidade de Brasília - UnB

Prof. Dr. Isaac de Lima Oliveira Filho
Membro Interno - Universidade do Estado do Rio Grande do Norte - UERN

Profa. Dra. Stela Mirla da Silva Felipe
Membro Externo - Universidade Estadual do Ceará - UECE

Agradecimentos

Agradeço primeiramente a Deus por ser um porto seguro nos momentos difíceis e, as minhas orientadoras Cicília Maia e Christina Pacheco que acreditaram no meu potencial e me deram todo o apoio necessário para a realização desse projeto. Vocês foram essenciais para a conclusão dessa etapa tão importante da minha vida e serei eternamente grato por tudo.

Agradeço também a minha esposa Aline Castro e minha filhinha Cecília que são incentivo e inspiração para a vida. Vocês foram essenciais para a superação dos diversos desafios e dificuldades impostas pela pandemia. Obrigado pela compreensão e apoio, sem vocês esse momento não seria tão especial.

Agradeço ainda a gestão e aos colegas de EEEP Professor Walquer Cavalcante Maia pela parceria e apoio. Sem vocês esse momento não seria possível.

Agradeço aos meus pais, minha irmã e todos os meus familiares que contribuíram direta e indiretamente para que eu chegasse até aqui.

Agradeço aos meus colegas de projeto e de mestrado que ajudaram nas mais diversas fases desse projeto e dos demais desenvolvidos durante o período do mestrado. Obrigado pela parceria, pela ajuda nos momentos de dificuldade, pela inspiração nos momentos de estudo.

Aos professores e demais funcionários da Universidade do Estado do Rio Grande do Norte (UERN) e da Universidade Federal Rural do Semi-Árido (UFERSA), obrigado por terem contribuído para o meu aperfeiçoamento profissional.

Por fim, obrigado a todos que contribuíram de alguma forma e incentivaram meu estudo. Obrigado por acreditar! Obrigado por apoiar! Cada um, de sua própria maneira foi extremamente importante para o que eu consegui construir até aqui.

Resumo

A conclusão do sequenciamento do genoma humano, em conjunto com o crescente avanço tecnológico possibilitou uma evolução significativa no âmbito da bioinformática, proporcionando o aprofundamento de diversos tipos de estudos e tecnologias de análise biológica, dentre estas, o microarranjo. O aumento de pesquisas utilizando essa tecnologia para análise de RNA gerou uma multiplicação na quantidade de dados disponíveis. A necessidade de publicar os dados brutos da pesquisa impulsionou a criação de bancos de dados públicos onde essas informações pudessem ser indexadas e resgatadas. Essas bases de dados são uma grande fonte de dados transcriptômicos que infelizmente acabam sendo subutilizadas. É comum que os softwares para reanálise desses dados não possuam uma interface gráfica e/ou apresentem lacunas em suas funcionalidades. Este trabalho objetiva o desenvolvimento de um sistema WEB para mineração de dados em estudos transcriptômicos a partir de microarranjos armazenados no banco de dados biológico GEO, esse sistema foi denominado de Mining_RNA. Para o desenvolvimento do sistema foi necessária uma revisão de literatura para listar as lacunas que pudessem ser exploradas por uma nova ferramenta. Esta etapa culminou no levantamento de requisitos e planejamento da arquitetura que seria seguida para o desenvolvimento do sistema, desenvolvido com as linguagens de programação PHP e Python para respectivamente visualizar e o processar os dados. Os resultados obtidos no sistema foram validados comparando-os com o sistema GEO2R, mantido pelo NCBI e confrontados também com os resultados originais do estudo analisado. O sistema Mining_RNA possibilita que, através de uma usabilidade passo-a-passo juntamente com uma série de filtros possam ser calculadas a expressão diferencial entre os genes de um estudo, possibilitando ainda a análise do cálculo do teste-t e do valor-p para cada gene do estudo analisado. A partir da validação dos dados gerados, foi possível perceber que o sistema teve uma eficácia aproximada de 98% na comparação com o GEO2R e de mais de 90% na comparação com o estudo original. O sistema desenvolvido pode ser um forte aliado dos pesquisadores para a reanálise de estudos biológicos, possibilitando uma nova forma de analisar os dados e gerando resultados tão confiáveis quanto ferramentas já consolidadas.

Palavras-chave: Bioinformática, Microarranjo, GEO, Sistema WEB, Mineração de Dados.

Abstract

The completion of the human genome sequencing, together with the technological advances, made a significant evolution in the scope of bioinformatics possible, providing a more in-depth types of studies and technologies of biological analysis, among them, the microarray. The increase in research using this technology for RNA analysis has generated a vast growth in the amount of available data. The need to publish research raw data boosted the creation of public databases where this information could be indexed and retrieved. These databases are a great source of transcriptomic data that unfortunately end up being underutilized. Softwares for reanalysis of this data generally lack graphical interfaces and/or have lacunae in its functionalities. Our aim is to develop of a WEB system for data mining in transcriptomic studies from microarrays stored in the GEO biological database, the system was named Mining_RNA. For the development of the system, it was necessary a literature review to list the gaps that could be explored by a new tool. This step culminated in the survey of requirements and planning of the architecture that would be followed for the development of the system, developed with the PHP and Python programming languages to visualize and process data respectively. The results obtained in the system were validated by comparing them with the GEO2R system, maintained by the NCBI and also compared with the original results of the analyzed study. The Mining_RNA system allows that, through a step-by-step usability together with a series of filters, the differential expression between the genes of a study be calculated, also allowing the analysis of the calculation of the t-test and the p-value for each analyzed gene. From the validation of the generated data, it was possible to notice that the system had an approximate efficiency of 98% in comparison with GEO2R and more than 90% in comparison with the original study. The developed system can be a strong ally of the researchers for the reanalysis of biological studies, enabling a new way to analyze the data and generating results as reliable as tools already consolidated.

Keywords: Bioinformatics, Microarray, GEO, Web System, Data Mining.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACM	Association for Computing Machinery
AE	Array Express
API	Application Programming Interface
BD	Banco de Dados
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CDNA	Complementary DNA
DNA	DeoxyriboNucleic acid
EBI	European Bioinformatics Institute
FC	Fold Change
GDS	GEO Data Set
GEA	Genomic Expression Archive
GEO	Gene Expression Omnibus
GPL	GEO Plataform
GSE	GEO Series
GSM	GEO Samples
IEEE	Institute of Electrical and Electronic Engineers
K-NN	K-Nearest Neighbors
LOG	logaritmo
MIAME	Minimum Information About a Microarray Experiment
NCBI	National Center For Biotechnology Information
OBJ	Objetivo
PGH	Projeto Genoma Humano
QC	Questão de Conhecimento

QP	Questão de Projeto
RF	Requisito Funcional
RNA	Ribonucleic Acid
RNA-SEQ	RNA Sequencing
RNF	Requisito Não Funcional
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SVM	Support Vector Machines
WEB	World Wide Web

Sumário

1 Introdução

1.1 Contexto

O projeto Genoma Humano teve a finalidade de identificar os genes existentes nos cromossomos através do sequenciamento do genoma humano. O projeto teve sua conclusão em 2003 (WONG, 2019), possibilitando que diversos estudos pudessem ser realizados objetivando a cura de diversas doenças, dentre estas o diabetes. A todo momento, o ser humano passa por transformações biológicas que definem o funcionamento do corpo, dentre as ações que acontecem continuamente está o processo de expressão gênica.

A expressão gênica consiste na produção de uma molécula funcional, ácido ribonucleico (RNA, do inglês Ribonucleic Acid) ou proteína, a partir da informação contida no ácido desoxirribonucleico (DNA, do inglês Deoxyribonucleic acid). Esse processo ocorre em duas etapas: transcrição e tradução. Na transcrição, o RNA é sintetizado a partir do molde de DNA, e o RNA produzido poderá ser posteriormente exportado para o citoplasma e usado para a sintetização de proteínas nos ribossomos. A transcrição é um dos processos onde a célula regula sua expressão gênica (ALVES; SOUZA *et al.*, 2013). A sintetização de proteínas a partir do RNA é chamada de tradução (PIERCE, 2016).

Os microarranjos de DNA estão entre as tecnologias em grande evidencia no campo da pesquisa genética (HERNÁNDEZ-CABRONERO *et al.*, 2016). O microarranjo de DNA permitiu o acúmulo de uma vasta quantidade de dados nas últimas décadas, transformando estudos biológicos de genes específicos a nível transcriptômico, impulsionando consideravelmente diversos campos dos estudos biológicos (SUN; SHAO; WANG, 2018).

Um estudo transcriptômico tem como objetivo isolar e caracterizar o RNA. Esse tipo de pesquisa consegue identificar informações a partir das sondas utilizadas pelo microarranjo e com isso prover uma enorme quantidade de dados para a pesquisa que está sendo realizada uma vez que, cada análise pode ter até 450.000 variáveis (sondas genéticas), fazendo com que, a depender do que está sendo pesquisado, a análise desses dados passe a ser bem complexa do ponto de vista computacional (HIRA; GILLIES, 2015).

Com o aumento de pesquisas utilizando microarranjo e a enorme quantidade de dados proveniente desses estudos, tornou-se necessária a criação de repositórios onde esses dados pudessem ser disponibilizados, a fim de confirmar as informações publicadas ou aprofundar a pesquisa, podendo ainda a partir de dados de mais de um estudo, realizar a inferência de novas informações (PAPATHEODOROU *et al.*, 2017). Os dados de expressão gênica identificados, são armazenados como microarranjo e *RNA Sequencing* (RNA-seq)

em três bancos de dados públicos, o Gene Expression Omnibus[1] (GEO), o ArrayExpress[2] (AE) e o Genomic Expression Archive[3] (GEA).

Outra característica importante para que esses estudos pudessem ter continuidade e validade foi a definição do padrão *Minimum Information About a Microarray Experiment* (MIAME). Este padrão descreve as informações mínimas em relação a um experimento de microarranjo de forma que seja possível a interpretação dos resultados do experimento de forma inequívoca e que possibilite a reprodução do experimento (BRAZMA, 2009).

As bases de dados de microarranjos são uma grande fonte de dados transcriptômicos que infelizmente acabam sendo subutilizadas (HUERTA *et al.*, 2014), no entanto, uma análise adequada dessas bases pode melhorar a nossa compreensão da biologia e da medicina (SCHMITZ-ABE *et al.*, 2019). Os estudos a partir dessas bases podem ocorrer de forma manual, no entanto é possível que, com o uso da tecnologia, essa análise possa tornar-se um processo mais eficiente, de forma que o cientista tenha seu foco nos dados, e não na carga de trabalho necessária para analisá-los.

1.2 Objetivos

Em virtude da metodologia *Design Science* para a produção dessa pesquisa, os objetivos serão divididos em objetivos dos *stakeholders* (partes interessadas no projeto) a qual seguirá a taxonomia descrita em Alexander (2005) e objetivos de pesquisa conforme proposto por Wieringa (2014).

1.2.1 Objetivos dos *stakeholders*

Os *stakeholders* explícitos nessa pesquisa são: os operadores normais, que nessa pesquisa são pesquisadores de áreas como: Biomedicina, Bioinformática, Genética, Microbiologia e áreas afins; O financiador é a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Os consultores, são profissionais da área da genética; e os Desenvolvedores, programadores responsáveis pela implementação do sistema desenvolvido.

Estes *stakeholders* trabalharam em sinergia objetivando desenvolver um sistema de mineração de dados oriundos de estudos transcriptômicos, este, é capaz de avaliar as mudanças na expressão gênica a partir de dados do repositório GEO para assim, ajudar os operadores normais a realizar novas pesquisas a partir dos dados obtidos.

[1] https://www.ncbi.nlm.nih.gov/geo
[2] https://www.ebi.ac.uk/arrayexpress/
[3] https://www.ddbj.nig.ac.jp/gea/index-e.html

1.2.2 Objetivos da pesquisa

- Realizar um levantamento dos projetos similares existentes na literatura;
- Obter e identificar dados de estudos transcriptômicos a partir da plataforma GEO;
- Construir um sistema WEB para que os operadores consigam realizar suas mineração e análises de dados;
- Desenvolver um algoritmo para avaliar mudanças na expressão gênica e integrá-lo no sistema WEB desenvolvido.

1.3 Questões de pesquisa

Os conjuntos de dados armazenados no repositório de informações biológicas GEO aumentam de forma acelerada. No entanto, ainda há uma dificuldade para realizar estudos nos metadados da pesquisa. Isso ocorre devido a falta da padronização no vocabulário empregado pelo autor da pesquisa para estruturar os dados que serão armazenados nas bases de dados (WANG; LACHMANN; MA'AYAN, 2019).

Outra vertente a ser analisada é relacionada ao conhecimento em linguagem de programação. Esse tipo de conteúdo não é comum em cursos nas áreas da saúde ou ciências biológicas, fazendo com que o uso de linguagens como o R, muito utilizada nas ferramentas voltadas para estudos biológicos, provoque dificuldade para quem deseja utilizar a bioinformática para aprofundar estudos. A partir disso é possível definir o seguinte questionamento: Um sistema que fizesse a leitura, pré-processamento, mineração, e exibição desses dados em uma interface de fácil acesso ajudaria pesquisadores a aprofundar ainda mais pesquisas já existentes?

Para tratar o questionamento, e seguindo a metodologia *Design Science* proposta por Wieringa (2014), esta pesquisa buscou responder as seguintes questões de conhecimento e de projeto. Ainda no contexto de *Design Science*, serão abordadas questões de *Trade-Off* e Satisfação.

- **Questões de Conhecimento**
  - **QC01:** Quais sistemas são utilizados atualmente para interpretação de micro-arranjos?
  - **QC02:** Existem maneiras de juntar os dados de mais de uma pesquisa a fim de buscar novos resultados?
  - **QC03:** Quais dados devem ser filtrados para que uma interface objetiva e relevante seja gerada?

- **Questões de Design/Projeto**
  - **QP01:** É possível utilizar a linguagem de programação PHP para o tratamento desses dados?
  - **QP02:** Que opções de filtragem devem ser disponibilizadas ao usuário a fim de ampliar suas possibilidades de resultados?
  - **QP03:** A utilização de um banco de dados relacional seria eficiente para guardar esses tipos de informações uma vez que, dependendo da pesquisa científica realizada, o conjunto de dados retornado pode ter diferenças significativas?

1.4 Motivação

A tecnologia da informação tornou-se essencial em diversos contextos cotidianos. Com isso, o uso de ferramentas que propiciam um auxílio de produtividade se tornou peça chave para conseguir melhores resultados nas mais diversas áreas de conhecimento (PIETKA *et al.*, 2019). Na informática médica, os sistemas de apoio a tomada de decisão podem auxiliar o profissional a evitar diagnósticos errados ou viciados, bem como, em tornar o processo mais ágil e objetivo (BENNETT; HARDIKER, 2016).

Para as ciências biológicas, o uso da bioinformática contribuiu para que o avanço na área e a produção de pesquisas pudessem caminhar aceleradamente. Hoje a bioinformática auxilia no processamento de grandes quantidades de dados (GASPAROVICA-ASĪTE; ALEKSEJEVA, 2019), podendo ainda realizar a mineração desses dados objetivando a extração de informações relevantes que possam auxiliar no tratamento ou detecção de enfermidades (GUDENAS *et al.*, 2019).

Embora a bioinformática seja uma área que remete as primeiras décadas da computação moderna, ainda existem muitas lacunas a espera de soluções eficientes. Uma dessas lacunas é a reutilização dos dados de pesquisas existentes de forma facilitada. Este projeto pretende implementar um sistema capaz de prover um serviço que recupere dados relacionados a pesquisas transcriptômicas a partir da plataforma GEO através de uma interface intuitiva. Esse serviço deve ser implementado na forma de um sistema WEB que possa auxiliar na descoberta de novas informações em pesquisas que, por dificuldade na análise dos dados encontram-se estagnadas.

Para um melhor entendimento dos recursos que serão desenvolvidos para suprir a carência detectada, a Tabela 1 associa os *stakeholders* a cada objetivo, ao artefato e aos Requisitos Funcionais (RF). Objetivando ainda uma melhor visualização da tabela os *stakeholders* serão divididos em Grupo 1 (G1) que compreenderá: Financiador, Consultores e Desenvolvedores, Grupo 2 (G2) que abrange: Operadores normais, Financiador, Consultores e Desenvolvedores.

Tabela 1 – Definição de requisitos funcionais por objetivo

Stakeholders	**Objetivos**	**Artefato**	**RF**
G1	**01:** Obter dados do estudo e armazenar em base de dados própria	Mining_RNA	**01:** Permitir a importação dos dados através do código da pesquisa
		Mining_RNA	**02:** Realizar a leitura dos dados e converte-los em consultas SQL
		Mining_RNA	**03:** Armazenar dados localmente em BD Relacional
G1	**02:** Desenvolver um algoritmo para avaliar mudanças na expressão gênica	Mining_RNA	**04:** Testar algoritmos de mineração de dados a fim de selecionar um ou mais que consiga extrair dados relevantes
		Mining_RNA	**05:** Realizar o cálculo de *Fold Change* de forma que possam ser observados os genes que diferem entre si considerando os diferentes grupos amostrais
G2	**03:** Construir um sistema WEB para que os operadores consigam realizar análises de dados	Mining_RNA	**06:** Possibilitar o uso de filtros de genes, amostras, grupos, dados estatísticos e, personalizados
		Mining_RNA	**07:** Possibilitar a operação em modo avançado ou em modo passo a passo
		Mining_RNA	**08:** Disponibilizar a opção de salvar os dados obtidos para continuação posterior da análise
		Mining_RNA	**09:** Possibilitar o compartilhamento dos resultados obtidos na análise

Fonte: Autoria Própria

Em conformidade com os dados apresentados na Tabela 1 foram definidos alguns

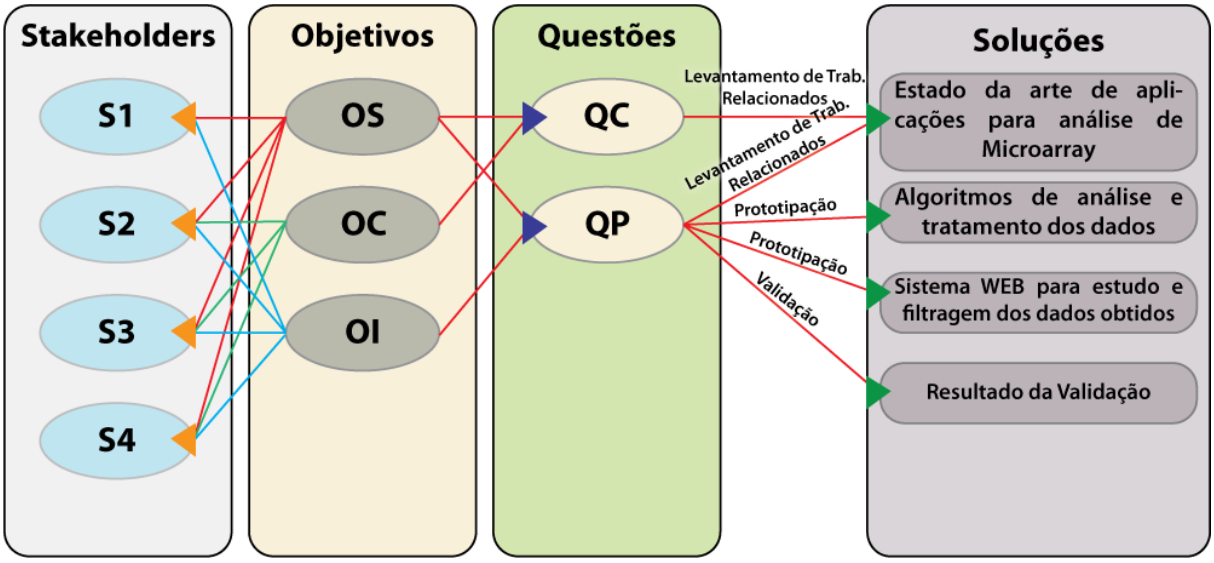
Requisitos Não Funcionais (RNF) necessários para o bom funcionamento do sistema proposto. Os RNF definidos foram:

- **RNF-01:** O sistema deve ser hospedado em nuvem;
- **RNF-02:** O sistema deve ser operacionalizado no sistema Linux;
- **RNF-03:** O uso de memória não deve ultrapassar 256MB por instancia executada;
- **RNF-04:** A interface do sistema deve ser otimizada para renderização em telas com no mínimo resolução 1024px x 768px.

1.5 Metodologia

Para ilustrar os métodos que foram aplicados para que os objetivos desse trabalho fossem atingidos, a Figura 1 demonstra o conjunto *stakeholders* e os objetivos pretendidos pelos mesmos. O usuário (S1) espera que o objetivo principal (OS) seja alcançado, esse também é o objetivo do financiador (S2), dos consultores (S3) e dos desenvolvedores que atuarão nessa pesquisa (S4). Os Objetivos de Conhecimento (OC) e de Instrumentação (OI) são almejados pelos *stakeholders* S2, S3 e S4. Ainda de acordo com a Figura 1 e considerando a coluna Questões, é possível perceber que o conjunto de objetivos OS e OC resultaram em Questões de Conhecimento (QC), já o conjunto de objetivos compreendido por OS e OI resultaram em questões de pesquisa (QP). Também fazem parte da referida figura o conjunto de soluções, que são os artefatos gerados pelas respostas de cada questão, juntamente com os métodos utilizados.

Figura 1 – Estrutura da pesquisa conforme o *Design Science*



Fonte: Autoria Própria

Para que as questões de pesquisa se tornassem artefatos relevantes para esta pesquisa foram utilizados os seguintes métodos:

- O levantamento de trabalhos relacionados teve papel bastante significativo para a realização deste trabalho acadêmico. Através dele pôde ser caracterizado o estado da arte relacionado ao projeto que foi implementado, a partir disto também foi possível identificar possíveis lacunas que pudessem ser preenchidas por um novo projeto. Algumas destas foram preenchidas pelo sistema implementado durante esta pesquisa. Para esse levantamento foram realizadas buscas avançadas em bases de dados conceituadas como, Science Direct, Elsevier, ACM e IEEE.
- No processo de prototipação foram utilizados *frameworks* de programação e de design, possibilitando um ganho de tempo no desenvolvimento. A partir disso foi possível desenvolver os algoritmos necessários e as interfaces que foram utilizadas no artefato.
- Na etapa de validação foram aplicadas as seguintes metodologias propostas por Wieringa (2014): Testes de software e Aplicação em caso único. Também fez parte desse processo, a realização da validação dos resultados conforme a metodologia MIAME proposta por Brazma (2009) a fim de comprovar se os resultados gerados pelo sistema atingem os requisitos necessários para a realização de novas publicações.

1.6 Organização do documento

Este trabalho está organizado da seguinte forma:

- O Capítulo 2 traz o referencial teórico e os trabalhos relacionados abordando: bioinformática, microarranjo, mineração de dados e a apresentação de trabalhos relacionados a mineração de dados de microarranjo.
- O Capítulo 3 demonstra o sistema proposto e sua arquitetura.
- No Capítulo 4 são apresentados os resultados obtidos durante a pesquisa e o desenvolvimento do projeto.
- Por fim, no Capítulo 5 são expostas as considerações finais a respeito deste trabalho, abordando ainda as perspectivas futuras para esse projeto.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os conceitos relevantes para a execução da pesquisa e o desenvolvimento da aplicação. Primeiramente é descrita a Bioinformática, na sequência são abordados conceitos sobre microarranjo além de conceitos relacionados a Mineração de Dados. E por fim, são descritos trabalhos relacionados a proposta desta pesquisa.

2.1 Bioinformática

A bioinformática é uma ciência que propõe a junção da Biologia e da Ciência da Computação para realizar a análise de grandes volumes de dados biológicos como sequências genéticas, amostras proteicas, dados celulares, entre outros, a fim de realizar previsões ou ainda realizar novas descobertas no campo da biologia (BAXEVANIS; BADER; WISHART, 2020). Na segunda metade do século XX, a informática e a biologia se desenvolviam de forma separada. No entanto, percebeu-se que com o uso de técnicas computacionais, seria possível analisar uma quantidade gigantesca de dados, proporcionando avanços significativos aos estudos biológicos.

O termo bioinformática foi utilizado inicialmente por Paulien Hogeweg e Ben Hesper em meados de 1970 para estudar processos informáticos em sistemas bióticos (GELDER *et al.*, 2019), sendo que somente a partir do final da década de 1980. O termo passou a ser utilizado essencialmente para se referir a métodos computacionais para análise comparativa de dados biológicos podendo ser aplicada às ciências ômicas como genômica, transcriptômica, proteômica, interatômica e metabolômica (ESPINDOLA *et al.*, 2010). A princípio, segundo os autores, o grande desafio era o poder computacional da época, então eles tentaram inicialmente o desenvolvimento de um conjunto integrado de métodos de análise de padrões supervisionados e não supervisionados, embora segundo os mesmos esse não fosse o real objetivo da bioinformática (HOGEWEG, 2011).

A evolução dos estudos biológicos possibilitou a obtenção de dados mais complexos, resultando em um aumento exponencial de informações biológicas, dessa forma, a análise manual desse conjunto de informações tornou-se muito difícil.Para auxiliar na resolução desse problema, foram empregados recursos da ciência da computação, com isso, os dados obtidos nos estudos puderam ser analisados com maior eficiência e precisão por meio de processos automatizados. Um dos exemplos de estudos complexos onde a bioinformática teve papel significativo foi o Projeto Genoma Humano (PGH), um estudo internacional de pesquisa científica concluído em abril de 2003 que disponibilizou pela primeira vez a sequência genética completa de um ser humano (BANAGANAPALLI; SHAIK, 2019).

Desde a conclusão do PGH, a geração de dados biológicos aumentou significativamente, e o domínio da bioinformática está desempenhando um papel fundamental na análise desses dados (GREENE; TROYANSKAYA, 2011). O PGH trouxe como uma de suas principais conquistas um melhor entendimento da dinâmica e modalidades dos transcritos do gene humano, a sua apresentação e localização no genoma e suas funções moleculares fundamentais. Com tal quantidade de dados, a geração de conhecimento dependia da mineração de dados eficiente. Dessa forma, o domínio interdisciplinar de bioinformática tem desempenhado um papel fundamental na mineração desses dados (BANAGANAPALLI; SHAIK, 2019).

2.2 Microarranjo

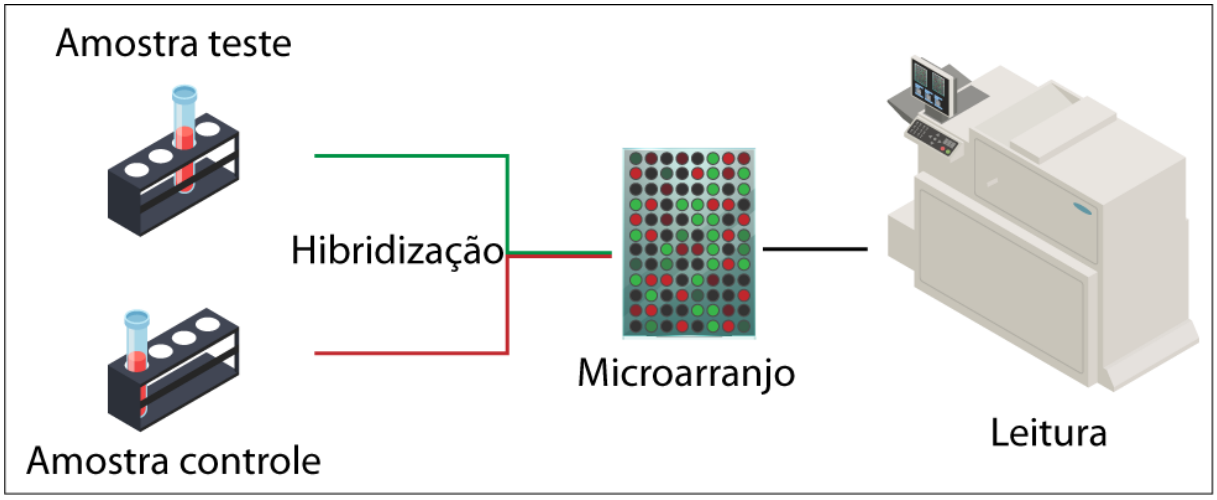
O microarranjo é uma tecnologia que permite quantificar simultaneamente a expressão de milhares de genes. Seu desenvolvimento remete aos anos 70, uma vez que seu funcionamento tem relação com uma técnica chamada *Dot-Blot* a qual foi desenvolvida naquele período, no entanto, somente durante a década de 90 a técnica ganhou as características utilizadas em seu funcionamento nos dias atuais. Para que isso fosse possível foi essencial a possibilidade de desenvolver arranjos de alta densidade a partir de sistemas robóticos automatizados, bem como o uso de métodos de detecção óptica para realizar a leitura de cada ponto da matriz (ROSA; ROCHA; FURLAN, 2007).

Para a aplicação da técnica, são necessárias placas de microarranjo, e um equipamento para a leitura dos dados conforme exemplificado na Figura 2. As placas são constituídas por lâminas de vidro, chips de silicone ou membranas de nylon. O elemento sólido escolhido é revestido com até milhares de sequências curtas de DNA (MISSOUM, 2018), divididos em pontos, também denominados de sondas. Estes pontos estão organizados em diversas linhas e colunas, assim como em uma matriz. Cada uma das sondas contém várias cadeias idênticas de DNA sendo que, a sequência de DNA em cada ponto é única, representando assim um gene (GONZALO; SÁNCHEZ, 2018).

Se considerado o uso de técnicas como *Northern blot* e *Reverse Transcription Polymerase Chain Reaction* (RT-PCR) que permitem testar apenas alguns genes por experimento, é possível afirmar que a quantidade de dados gerados em um teste com microarranjo ou perfil de expressão global é muito maior, uma vez que esse tipo de experimento analisa uma quantidade maior de genes se comparado com os outros testes citados (GOVINDARAJAN *et al.*, 2012).

Os microarranjos de DNA passaram a ser uma das ferramentas mais amplamente utilizadas na última década em biologia molecular (WORKU; NEGASSU, 2019) uma vez que, é uma forma de pesquisa que pode ser utilizada para uma série de experimentos, incluindo: para a determinação de perfis de expressão gênica, para o estudo de genômica

Figura 2 – Processo de análise de microarranjo



Fonte: Autoria Própria

funcional (GUINDALINI; TUFIK, 2007), na análise de *splicing* alternativo, na análise da hibridização genômica comparativa a fim de identificar amplificações genéticas e deleções (DORFMAN *et al.*, 2015), entre outras utilizações.

Em análises ligadas ao fenótipo, é possível afirmar que até pequenas variações na sequência de DNA que levam a diferentes características (como cor da pele, características faciais ou altura), conhecidas como polimorfismos, que podem causar ou contribuir para o desenvolvimento de muitas síndromes e doenças, também podem ser facilmente identificados pela técnica de microarranjo (GOVINDARAJAN *et al.*, 2012).

Os estudos realizados pela comunidade científica com o auxílio da tecnologia de microarranjos geraram uma enorme quantidade de dados nos últimos anos, e seria inviável a publicação desses dados brutos juntamente com a publicação gerada a partir pesquisa uma vez que são séries de dados extensas. Para facilitar a acessibilidade a esses dados, o *National Center For Biotechnology Information* (NCBI) formulou o *Gene Expression Omnibus* ou GEO, já o *European Bioinformatics Institute* (EBI) mantém o *ArrayExpress* (ROSA; ROCHA; FURLAN, 2007). Ambos os bancos de dados são públicos e contém dados brutos de diversas pesquisas, possibilitando que a partir desses dados, outros pesquisadores possam confirmar os resultados dos estudos, bem como realizar novas descobertas a partir da releitura dessas informações.

2.3 Mineração de dados

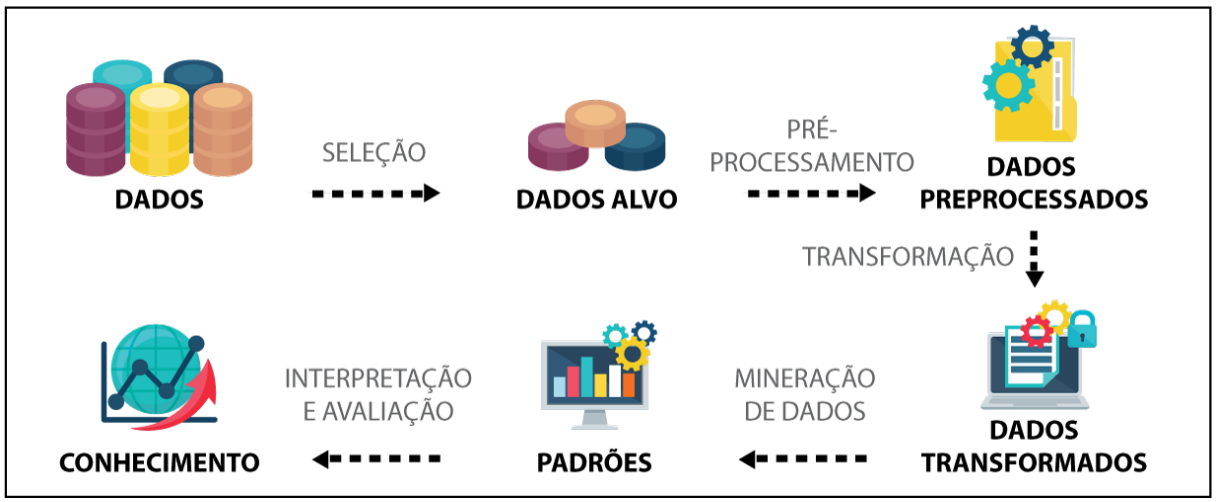
A mineração de dados é o processo de coleta, assimilação e utilização de informações objetivando a descoberta de anomalias ou de padrões consistentes, que representem algum ganho de conhecimento. Os dados geralmente são coletados de grandes bancos de dados, também conhecidos como *big data*, e processados para determinar padrões e outras

correlações (SEDKAOUI, 2018).

Esses padrões podem ser estatísticos, por exemplo, a previsão e derivação da taxa de desemprego a partir da mineração de dados. As correlações também podem ser usadas no domínio do aprendizado de máquina, como no caso da incorporação de técnicas de mineração em softwares que podem ser utilizados para prever o comportamento de clientes (SHMUELI *et al.*, 2019). A mineração de dados também é conhecida como *Knowledge Discovery in Databases* (KDD), esse processo consiste em extrair informações anteriormente desconhecidas, mas potencialmente úteis a partir de bases de dados. O processo de KDD consiste na execução de uma série de etapas com objetivo de transformar dados brutos em conhecimento relevante (PIATESKI; FRAWLEY, 1991). Essas etapas são demonstradas pela Figura 3 e descritas a seguir:

- **Seleção:** é definido o conjunto de dados alvo nos quais é desejado que seja realizada a mineração;
- **Pré-processamento:** tem como finalidade a limpeza de inconsistências ou informações ausentes, realizar a combinações quando houver o uso de múltiplas bases de dados e tratar anomalias relacionadas ao conjunto de informações analisadas, também conhecidas como *outliers*. Esse tratamento é executado visando que o resultado da mineração não seja comprometido por problemas nas entradas;
- **Transformação:** os dados são reorganizados e consolidados para que seja dado seguimento ao processo de mineração;
- **Mineração de dados:** nesta etapa são aplicados métodos inteligentes para que padrões interessantes sejam extraídos;
- **Interpretação e avaliação:** é feita interpretação dos dados para identificar se os padrões identificados são de interesse e representam algum conhecimento relevante. Ainda nesta etapa são aplicadas técnicas de visualização e representação para que a informação extraída seja apresentada ao usuário.

Figura 3 – Etapas do processo de KDD



Fonte: Autoria Própria

Técnicas de mineração de dados

Existem dois modelos principais para a mineração de dados, o preditivo e descritivo. E para cada um desses modelos existem algumas técnicas para as operações de mineração.

Técnicas preditivas de mineração de dados

A análise preditiva de dados, como o próprio nome sugere, visa prever resultados com base em um conjunto de padrões e tendências (PEREIRA, 2014). As técnicas preditivas de mineração de dados mais comuns incluem regressão e classificação:

- **Regressão:** usada para prever um intervalo de valores numéricos, dado um conjunto de dados específico. Por exemplo, a regressão pode ser usada para prever o custo de um produto ou serviço, considerando para isso dados um conjunto de outras variáveis;
- **Classificação:** semelhante à regressão, a classificação tenta prever um resultado, no entanto sem ser necessariamente um valor numérico. Por exemplo, uma determinada empresa do ramo bancário utilizar-se de uma série de informações para prever qual a probabilidade de um indivíduo pagar um empréstimo. As classificações podem ser coisas como "ruim", "razoável", "boa" e "excelente".

Técnicas descritivas de mineração de dados

A análise descritiva dos dados se baseia em dados históricos para entender tendências e avaliar mudanças ao longo do tempo. As técnicas descritivas mais comuns de mineração de dados incluem regra de associação e agrupamento (clusterização):

- **Regra de associação:** esse tipo de mineração de dados procura associações (coisas como padrões e correlações) com base em grandes quantidades de informações extraídas dos bancos de dados. Um exemplo do uso desse método seria, a partir da análise dos produtos comprados por consumidores de um determinado supermercado, determinar sempre que o consumidor compra os produtos W, X e Y, há uma probabilidade que esse consumidor também compre o produto Z. Na literatura há um exemplo clássico que cita a correlação de fraldas e cerveja (CHEN; CHEN; TUNG, 2006);

- **Agrupamento:** é o processo de transformar um grupo de objetos abstratos em classes de objetos semelhantes tendo como base para isso determinadas características dos objetos. Como exemplo dessa técnica poderíamos citar a criação de uma lista de segmentação de marketing com base na análise de buscas de produtos de usuários, dessa forma os usuários não seriam mais vistos como um simples indivíduo com determinadas preferencias, mas sim como um elemento que faz parte de um grupo.

Classificadores e comitês

Um algoritmo desenvolvido objetivando a classificação de instâncias em uma determinada base de conhecimento é denominado classificador. Atualmente é comum encontrar novos classificadores. O desafio é que esses classificadores sejam precisos e eficientes uma vez que cada vez mais tarefas importantes em análise de dados são delegadas a esses algoritmos. Embora alguns algoritmos ainda careçam de testes mais profundos, algumas técnicas de mineração já são bastante conhecidas e descritas na literatura, tais como, árvores de decisão, redes neurais, *Support Vector Machines* (SVM), *k-Nearest Neighbors* (k-NN) e *Naïve Bayes*.

Os algoritmos utilizados na mineração ou na área de aprendizagem de máquina são classificados em quatro categorias:

- **Aprendizagem supervisionada:** os algoritmos são treinados usando exemplos rotulados, como uma entrada em que a saída desejada é conhecida. O algoritmo de aprendizado recebe um conjunto de entradas junto com as respectivas saídas corretas, com base nisso, a aprendizagem é realizada comparando saídas previstas com as saídas corretas afim de encontrar erros. Após o treinamento é possível com base nos atributos que a instância analisada possui, deduzir a qual classe de dados ela pertence. Com base nas métricas identificadas durante o treinamento, é possível ainda, identificar qual a acurácia daquele teste.

- **Aprendizagem semi-supervisionada:** é usado para as mesmas aplicações que o aprendizado supervisionado. Nesse tipo de classificação tanto os dados rotulados

quanto os não rotulados são explorados para realizar a aprendizagem de forma empírica, ao invés de realizar a classificação com base nos rótulos existentes (HUSSAIN; CAMBRIA, 2018).

- **Aprendizagem não supervisionada:** nesse tipo de análise, o processo de aprendizagem é denominado dessa forma pelos dados utilizados não possuírem rótulos previamente conhecidos, o próprio algoritmo será empregado para descobrir as classes dos dados. O sistema não recebe a resposta certa. O algoritmo deve descobrir o que está sendo mostrado de maneira exploratória e, a partir disso tentar encontrar alguma informação útil. Nessa categoria são comuns algoritmos de agrupamento. Esse processo pode ser ilustrado por um exemplo onde o algoritmo recebe um conjunto de imagens de dígitos manuscritos. Em um cenário onde o algoritmo encontre 10 conjuntos de dados distintos, ele realizará o agrupamento desses dados em *clusters*. Os 10 dígitos distintos identificados podem corresponder ao intervalo numérico de 0 a 10. Este modelo não diz o que significa semanticamente cada grupo, uma vez que os dados não são rotulados, mas desse agrupamento podem ser extraídas informações úteis (HAN; KAMBER; PEI, 2011).

- **Aprendizado por reforço:** frequentemente utilizado em robótica e jogos. Nesse tipo de aprendizagem, o algoritmo realiza suas descobertas com base na tentativa e erro. Nesse tipo de aprendizado, a questão é abordada como um agente autônomo que, com base em suas ações em um ambiente, passa a escolher as ações ideais para atingir seus objetivos. Durante o treinamento do agente, cada vez que o mesmo executa alguma ação em seu ambiente, um treinador pode conceder uma penalidade ou recompensa para indicar o quanto aquela ação é interessante (MITCHELL, 1997).

Os Métodos de Mineração de Dados do tipo Comitê, também conhecidos como *ensemble* ou Combinadores de Modelos, são métodos de aprendizado de máquina que aproveitam o poder de vários classificadores para obter melhor precisão de previsão do que qualquer um dos classificadores individuais poderia fazer por conta própria (Seni; Elder, 2010). O objetivo básico ao projetar um conjunto é o mesmo que ao estabelecer um comitê de pessoas: cada membro do comitê deve ser o mais competente possível, mas os membros devem ser complementares entre si. Se os membros não são complementares, ou seja, se sempre concordam, o comitê é desnecessário, podendo dessa forma ser consultado individualmente. Se os membros forem complementares, quando um ou alguns membros cometerem um erro, a probabilidade é alta de que os membros restantes possam corrigi-lo (OZA, 2019).

A mineração de dados biológicos

Com a quantidade de dados brutos gerados constantemente seria uma tarefa complexa o processamento desses dados manualmente (HIRA; GILLIES, 2015). Conseguir unir os dados brutos de uma análise, utilizando microarranjo de duas ou mais pesquisas, acarretaria em um esforço gigantesco por parte dos pesquisadores. Para facilitar o processo, o cientista pode fazer uso de técnicas de mineração de dados para conseguir extrair dados novos de pesquisas preexistentes.

Uma das premissas básicas adotadas pela bioinformática é a mineração de dados biológicos (GANGWAR; GHOSE; SINGH, 2012). A utilização da mineração de dados pode ajudar a transformar dados que não são facilmente compreensíveis em uma interpretação meramente visual, podendo ainda, facilitar a inferência de informações novas e úteis para a descoberta de novos resultados em pesquisas novas ou existentes (GARG; MAHAJAN; KAMAL, 2017). Os novos resultados em pesquisas existentes se dariam pela nova abordagem do pesquisador em relação aos dados previamente analisados em outra pesquisa. Muitas vezes, esses dados estavam subutilizados nos bancos de dados biológicos públicos, podendo com essa nova análise prover novos avanços em suas respectivas áreas de pesquisa (LAN *et al.*, 2018).

Atualmente, uma das formas possíveis para a realização de mineração de dados biológicos é o *Bioconductor*[1] (NIE *et al.*, 2009), um projeto baseado na premissa do software livre que tem como objetivo promover a análise estatística e a compreensão dos estudos biológicos de alto rendimento, novos e preexistentes. O projeto é baseado em pacotes escritos, em grande parte, utilizando a linguagem de programação R, podendo conter, no entanto, contribuições de outras linguagens. O *Bioconductor* é uma das mais relevantes ferramentas para estudo de dados biológicos. Os pacotes disponibilizados a partir de seu repositório são destinados a diversos tipos de análises, dentre elas, a mineração.

2.4 Sistemas de análise de dados de microarranjos

A disponibilização de dados brutos de estudos biológicos no GEO possibilita a reprodutibilidade de estudos e facilita a reutilização desses dados. Reanalisá-los pode levar a novos *insights* científicos. No entanto, para que essas análises possam ser replicadas é necessária uma aplicação mais rigorosa dos requisitos definidos na metodologia MIAME, bem como, o desenvolvimento e uso de ferramentas que possibilitem a análise desses dados de forma fácil (RUNG; BRAZMA, 2013). A partir do levantamento apresentado a seguir é possível perceber que existem lacunas que podem ser preenchidas por outros softwares de bioinformática. Os trabalhos encontrados na literatura utilizam funcionalidades como

[1] https://www.bioconductor.org/

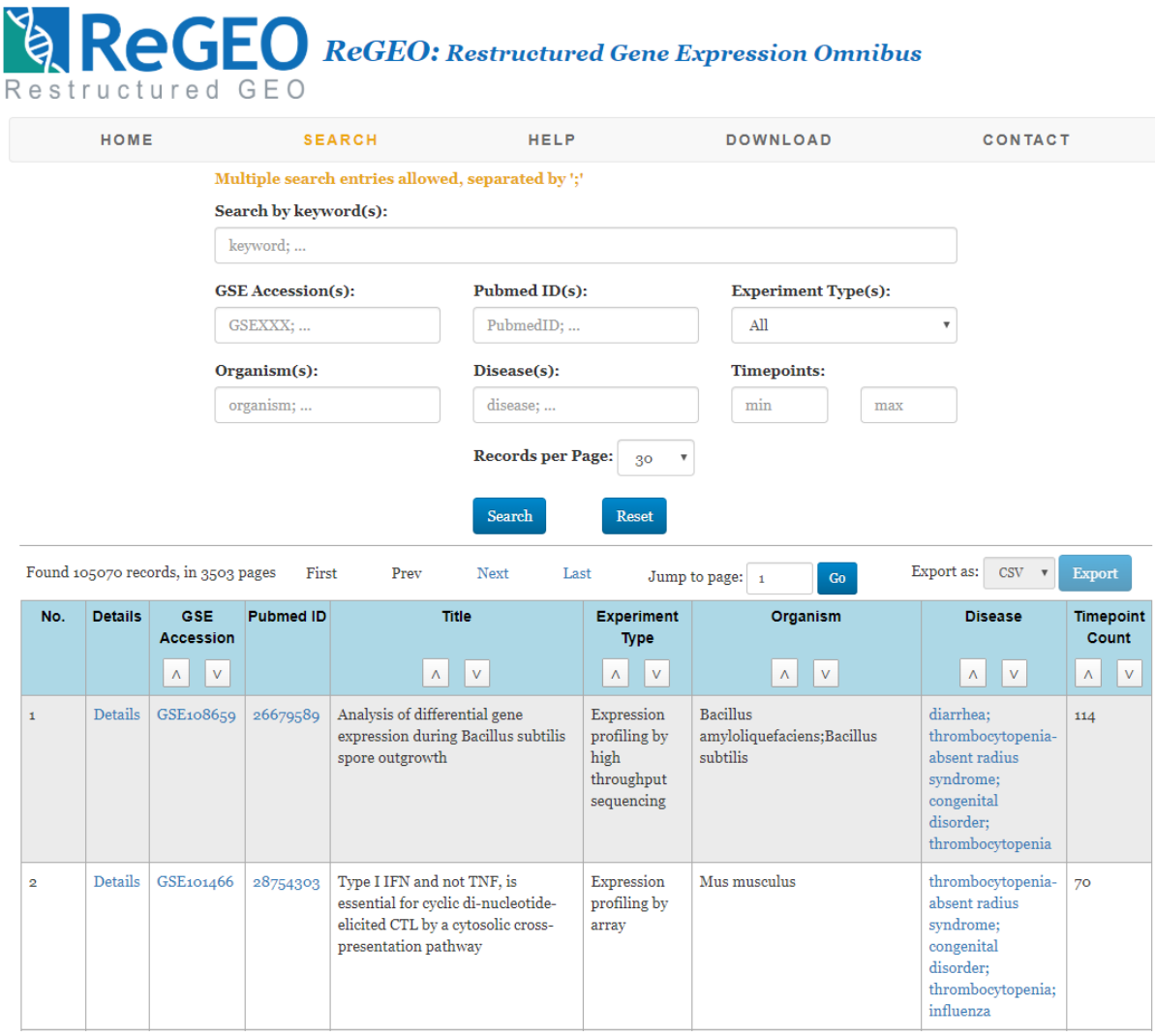
mineração de metadados, filtragem paramétrica, interface intuitiva, gráficos e, comparação múltiplas de pesquisas.

Davis e Meltzer (2007) desenvolveram um sistema denominado GEOquery. A ferramenta implementada nesse estudo é uma biblioteca desenvolvida na linguagem de programação R e disponibilizada gratuitamente através do repositório de ferramentas de bioinformática Bioconductor[2]. Disponibilizada em 2006 a ferramenta promete ser uma ponte entre a plataforma GEO e o Bioconductor. De fato, a ferramenta cumpre sua premissa básica e consegue captar os dados armazenados no GEO, bem como retornar através de seus métodos, as informações contidas no arquivo capturado. A biblioteca consegue capturar o arquivo que se deseja analisar a partir do código (exemplo GDS3884), copiando o arquivo para a máquina onde se está executando o comando e a partir disso, consegue trabalhar com os dados contidos no mesmo. A ferramenta em questão não possui uma interface gráfica e isso dificulta seu uso por parte de usuários sem conhecimento de programação.

Chen *et al.* (2019) observando que o GEO, bem como outros bancos de dados biológicos não armazenavam os dados dos estudos de forma estruturada, o que dificultava a reutilização e pesquisa desses dados, desenvolveram uma ferramenta a qual denominaram de *Restructured GEO* (ReGEO[3]). Ao analisar a interface implementada na ferramenta, é possível perceber que houve uma significativa melhora na interface conforme demonstra a Figura 4, da forma apresentada pelos autores, o ReGEO tem um ambiente bem mais amigável se comparado com o GEO.

[2] https://bioconductor.org/packages/release/bioc/html/GEOquery.html
[3] http://www.regeo.org

Figura 4 – Screenshot da Tela do ReGEO



Fonte: Adaptado de Chen *et al.* (2019)

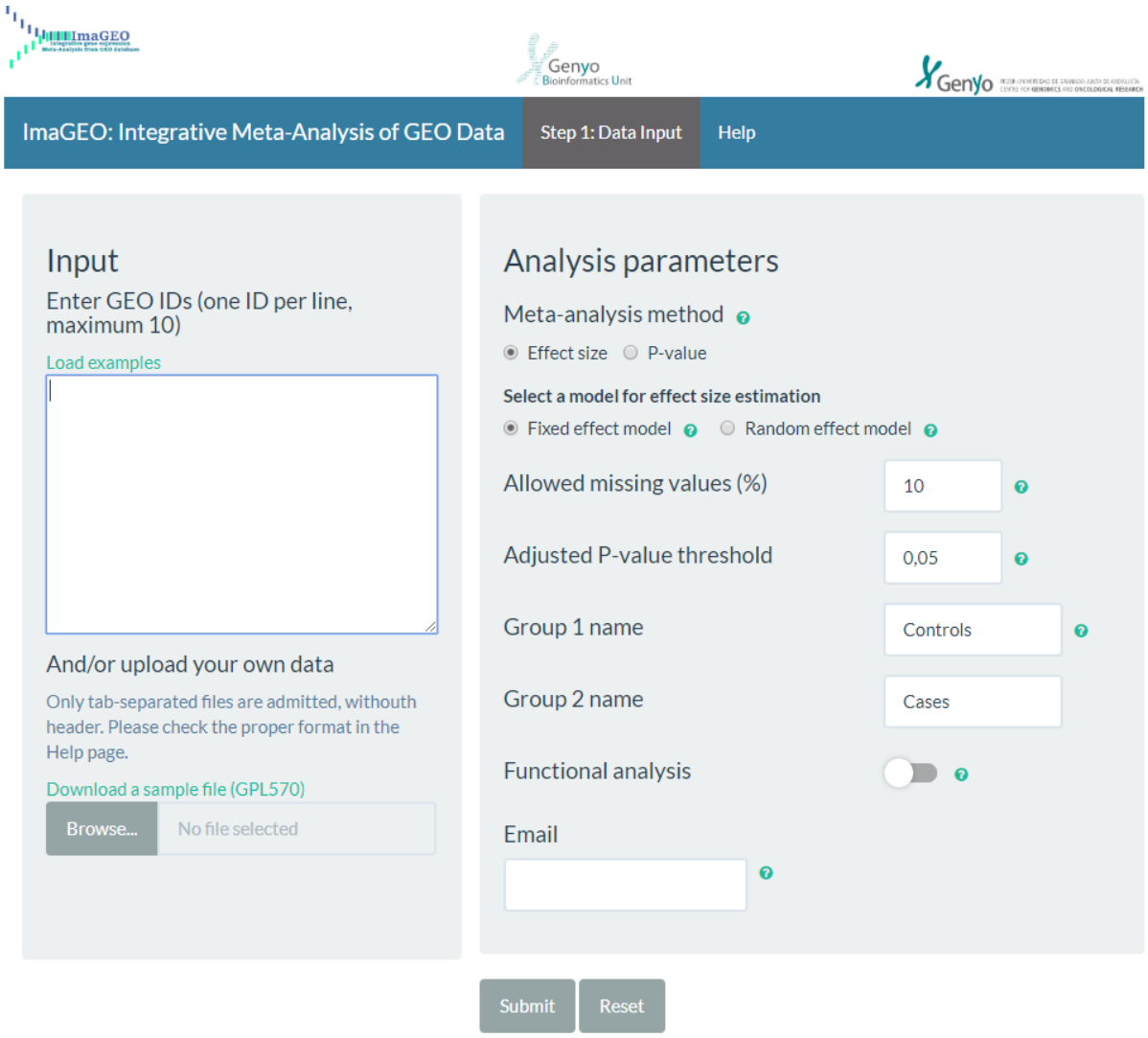
Para o desenvolvimento da ferramenta, os autores utilizaram técnicas de mineração de texto com o objetivo de analisar os metadados mantidos pelo GEO. Com isso, a ferramenta ReGEO reorganiza e categoriza as séries GEO tornando-as pesquisáveis por dois novos atributos extraídos automaticamente dos metadados de cada série, um desses atributos é a doença analisada, informação que não era indexada anteriormente. A busca por palavras chave também está disponível, assim como na ferramenta original. A abordagem utilizada pelos autores atingiu uma taxa de precisão na mineração de texto de 93,5% de forma totalmente automática.

Em sua pesquisa, Toro-Domínguez *et al.* (2018) apresentou o desenvolvimento da plataforma ImaGEO[4]. A ferramenta desenvolvida é baseada em um pacote chamado *Shiny*, uma biblioteca voltada a implementar uma interface WEB em conjunto com a linguagem de programação R. A solução apresentada conta com interessante sistema que divide-se

[4] http://bioinfo.genyo.es/imageo/

em 5 módulos. Inicialmente os dados são carregados e processados utilizando um conjunto de parâmetros que podem ser personalizados pelo usuário conforme exposto na Figura 5.

Figura 5 – Screenshot da Tela do ImaGEO

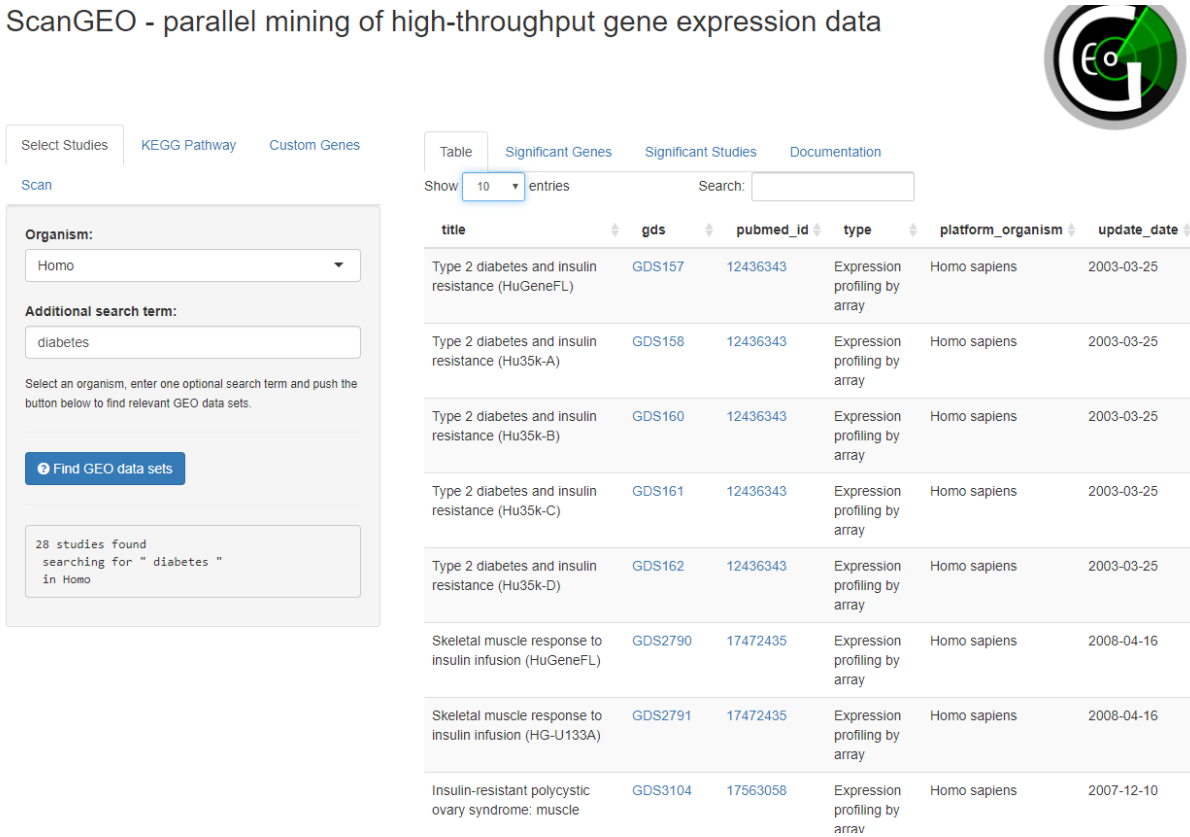


Fonte: Adaptado de Toro-Domínguez *et al.* (2018)

Para o carregamento dos dados, o pacote GEOquery desenvolvido por Davis e Meltzer (2007) é utilizado. Outros módulos implementados pela ferramenta são para: controle de qualidade, metanálise do *dataset*, análise funcional e por fim o módulo responsável pelo relatório. Este engloba a visão geral dos parâmetros utilizados na análise, um resumo relacionado ao processamento dos dados detectando por exemplo *outliers* e valores ausentes. Por fim, apresenta também os resultados do processamento de dados, incluindo até gráficos de calor objetivando uma melhor representação das informações obtidas sobre os genes. O ImaGEO é um recurso bastante completo e com interface bastante intuitiva, no entanto, ainda existem lacunas a serem preenchidas no quesito visualização final dos dados.

Outra aplicação com teor similar é o ScanGEO[5] desenvolvido por Koeppen, Stanton e Hampton (2017). A ferramenta é escrita em R e também é implementada utilizando o pacote *shiny*. O ScanGEO possibilita consultas aos dados do GEO identificando estudos relevantes conforme os parâmetros preenchidos pelo usuário. Segundo os autores a aplicação suporta a análise da matriz de expressão gênica dos 20 principais organismos presentes nas bases de dados do GEO, bem como, o uso de palavras chave para a busca conforme disposto na Figura 6. Ao realizar a pesquisa o usuário pode até realizar a exportação dos dados brutos obtidos para arquivos CSV, para assim, analisá-los localmente ou até em outros softwares relacionados. Ainda segundo o autor, a ferramenta pode acelerar consideravelmente a análise de dados públicos para a partir disso gerar hipóteses e realizar a validação de dados experimentais (KOEPPEN; STANTON; HAMPTON, 2017).

Figura 6 – Screenshot da Tela do ScanGEO



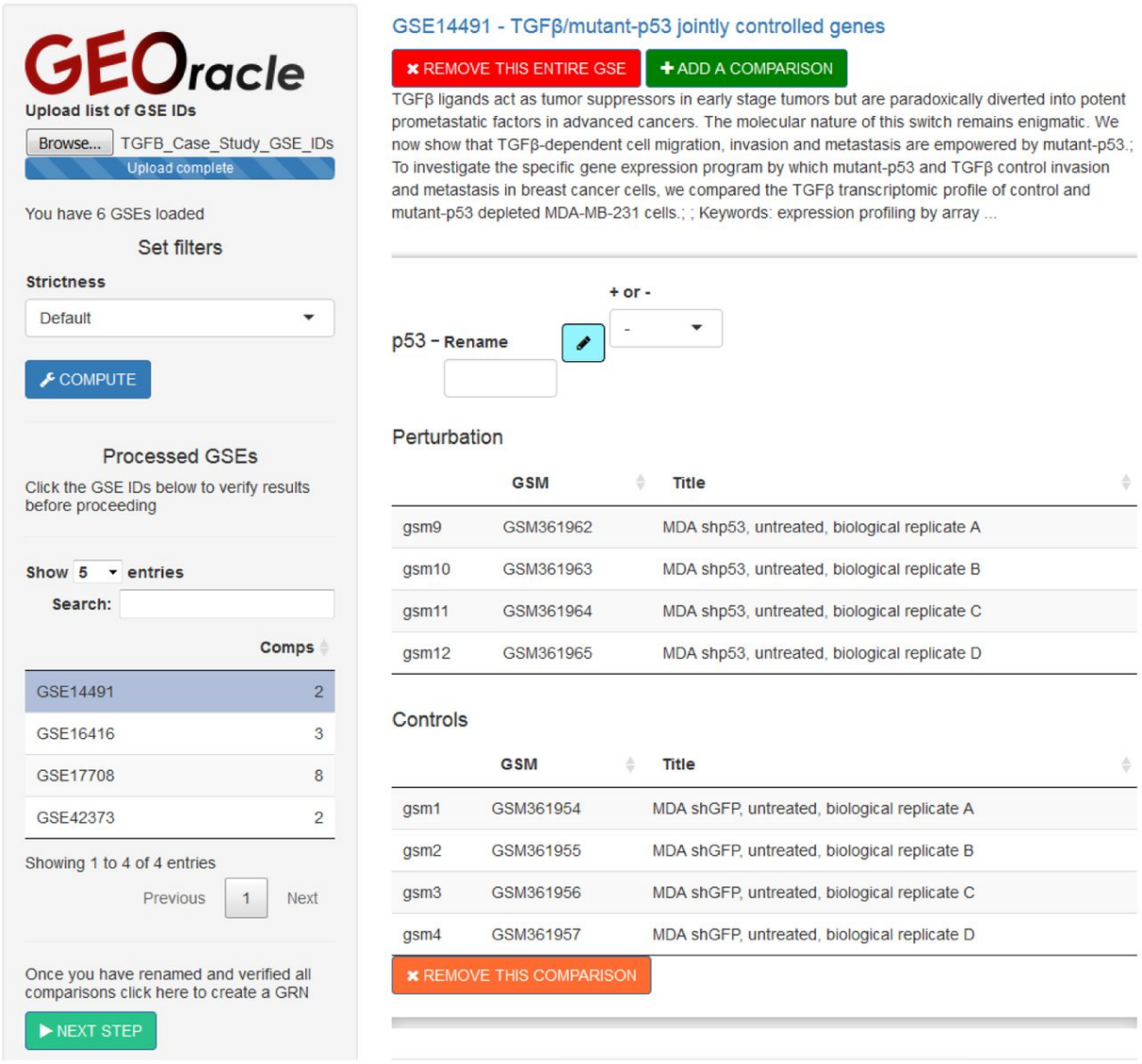
Fonte: Adaptado de Koeppen, Stanton e Hampton (2017)

Djordjevic *et al.* (2019) apresentam em seu artigo uma ferramenta capaz de utilizar técnicas de mineração de texto e aprendizado de máquina para identificar automaticamente experimentos de perturbação (ex. nocaute gênico), separa os diferentes grupos (experimental e controle) e avalia a expressão diferencial, o GEOracle. O sistema supracitado é uma ferramenta de código livre e embora o autor não informe um endereço onde a ferramenta pode ser executada diretamente, ele disponibiliza o código fonte armazenado na plataforma

[5] http://scangeo.dartmouth.edu/ScanGEO/

GitHub[6]. A interface da ferramenta é intuitiva, mas conta com menos funcionalidades se comparada com o ImaGEO conforme demonstra a Figura 7

Figura 7 – Screenshot da Tela do GEOracle



Fonte: Djordjevic *et al.* (2019)

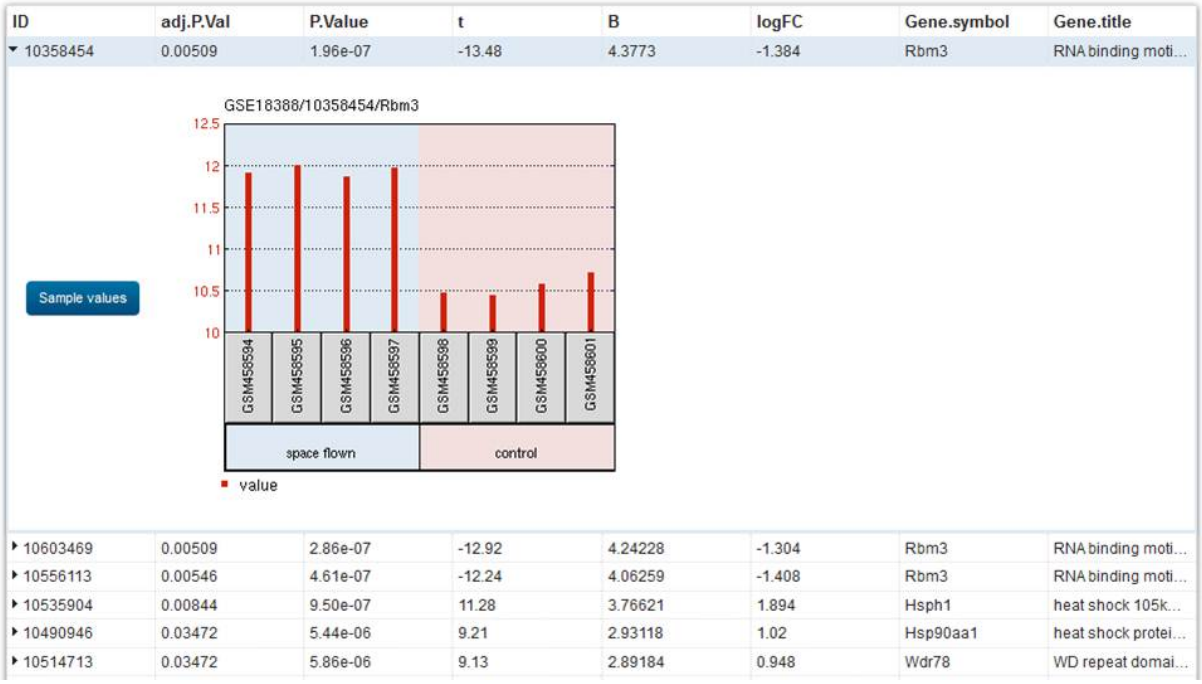
O GEO2R é uma ferramenta mantida pelo NCBI que permite aos usuários comparar dois ou mais grupos de amostras a partir de estudos armazenados no GEO. O sistema permite que seja identificada a expressão diferencial dos genes de um estudo, retornando os dados na forma de uma tabela com os genes ordenados por significância, além de retornar uma serie de gráficos para ajudar na visualização dos genes diferencialmente expressos (NCBI, 2021a).

Na interface online do GEO2R é possibilitado que o pesquisador analise os 250 genes mais relevantes, para visualizar o restante dos genes o sistema possibilita que o usuário baixe a tabela para que esta possa ser carregada em um outro programa como o

[6] https://github.com/VCCRI/GEOracle

Microsoft Excel ou similar. O sistema retorna ainda um código em R que o pesquisador pode executar no próprio computador, no entanto para conseguir alterar parâmetros é necessário o conhecimento na linguagem de programação. A Figura 8 é parte da interface do GEO2R, nela é possível visualizar parte da tabela com os dados de um estudo analisado.

Figura 8 – Screenshot da Tela do GEO2R



Fonte: NCBI (2021a)

Os trabalhos relacionados tem como principal objetivo facilitar a realização de análises de dados por parte dos pesquisadores. Em grande parte, são ferramentas que conseguem realizar a leitura de parte dos dados que a plataforma GEO não deixa exposta em sua interface, uma vez que algumas destas informações não estão devidamente alocadas, e sim descritas no contexto do estudo. Outras ferramentas conseguem fazer a ligação com informações de outras plataformas de pesquisa. Em algumas, os dados são processados instantaneamente, em outras, o resultado é enviado por e-mail ao final da análise.

Apesar de alguns trabalhos apresentarem funcionalidades de mineração em texto e aprendizagem de máquina, é possível aprimorar o processo de visualização dos dados obtidos a partir da disponibilização de uma série de filtros direcionados para obter resultados mais significativos a partir do conjunto estudado. Esses filtros devem incluir uma mineração nos dados brutos do microarranjo, análise de expressão diferencial e a devida filtragem dos dados de forma integrada. São essas funcionalidades que foram implementadas no presente trabalho. A Tabela 2 compara as funcionalidades disponibilizadas por cada ferramenta com as funcionalidades definidas para o sistema Mining_RNA, desenvolvido neste trabalho.

Tabela 2 – Comparativo entre os trabalhos relacionados e o sistema proposto

	GEO-query	Re-GEO	Ima-GEO	Scan GEO	GEO -racle	GEO2R	Mining_ RNA
Captura de dados a partir do GEO	X	X	X	X	X	X	X
Interface WEB obedecendo os critérios de usabilidade		X	X	X	X	X	X
Mineração de texto		X	X	X			X
Comparação entre grupos de um mesmo estudo		X	X	X	X	X	X
Cálculo de *Fold Change*			X			X	X
Filtragem na visualização dos dados relacionados a expressão gênica							X
Mineração de dados de expressão gênica			X				X

Fonte: Autoria Própria

3 MINING_RNA: Sistema WEB para mineração de dados em estudos transcriptômicos a partir de microarranjos

Este capítulo é destinado à apresentação do Mining_RNA, objeto de desenvolvimento desta dissertação de mestrado. Neste capítulo será apresentado o contexto no qual foi pautado o desenvolvimento, o funcionamento da aplicação, a organização da arquitetura e a metodologia aplicada durante o desenvolvimento.

3.1 Descrição e contexto do sistema desenvolvido

Nos últimos anos, diversos dados advindos de pesquisas científicas na área biológica foram disponibilizados nos bancos de dados públicos dedicados ao armazenamento e exposição dessas informações (BONO, 2020). Em intensidade bastante similar, vem surgindo nas últimas décadas ferramentas computacionais capazes de interpretar dados biológicos a fim de produzir novos resultados (WANG; LACHMANN; MA'AYAN, 2019).

Uma das bases de dados transcriptômicos com maior relevância na comunidade cientifica é a *Gene Expression Omnibus*. Sua plataforma conta com quase 4.500 estudos armazenados na forma de *datasets*, estudos esses que analisaram quase 3 milhões e meio de amostras para gerar seus resultados (WANG; LACHMANN; MA'AYAN, 2019). Atualmente a forma que a plataforma disponibiliza esses dados não é amigável a usuários sem conhecimento técnico apropriado para leitura de informações em arquivos de texto simples, tendo seus dados divididos em linha e separados por tabulações. A partir disso faz-se necessário o uso de ferramentas que possam auxiliar nessa dificuldade para que assim, o pesquisador possa dedicar seu esforço aos dados, e não a conseguir entender o modelo de armazenamento.

O presente trabalho teve como objetivo implementar um sistema capaz de prover um serviço que recupere os dados a partir da plataforma GEO. Esses dados devem ser pré-processados de forma que as informações oriundas da plataforma sejam validadas a fim de passar corretamente pelas etapas posteriores que podem incluir, de acordo com a preferência do usuário, a mineração desses dados ou ainda o processamento manual das informações a partir de filtros disponibilizados no sistema.

Durante o desenvolvimento do sistema Mining_RNA, bem como nas etapas posteriores a implementação do mesmo, objetivou-se alcançar os seguintes resultados:

1. Implementar uma Application Programming Interface (API) de interação entre o banco de dados biológico GEO e um banco de dados local;
2. Desenvolver uma interface para pré-processamento dos dados obtidos;
3. Implementar técnicas de mineração e aprendizagem de máquina para atuar em cima dos dados pré-processados;
4. Desenvolver uma interface análise e interação entre o pesquisador e os dados do estudo;
5. Possibilitar que pesquisadores sem conhecimento em programação possam utilizar os dados de bancos de dados biológicos;
6. Proporcionar a extração de novas informações de dados brutos disponibilizados publicamente no banco de dados biológico GEO;
7. Desenvolver um sistema que possa ser utilizada no cotidiano do pesquisador da área de bioinformática e de áreas afins.

No intuito alcançar os resultados supracitados foi desenvolvido um sistema WEB, baseado nas especificações levantadas a partir de estudo realizado junto a especialista de domínio que atua na área da biologia / bioinformática. O sistema suporta inicialmente a utilização de dois idiomas, o português e o inglês, este segundo será adotado objetivando facilitar o uso pela comunidade internacional, tendo em vista que está entre os idiomas mais utilizados no mundo, e figura-se como uma espécie de linguagem universal. O uso do sistema segue o estilo passo-a-passo onde o pesquisador avança entre as telas até chegar na listagem dos dados filtrados durante as etapas anteriores.

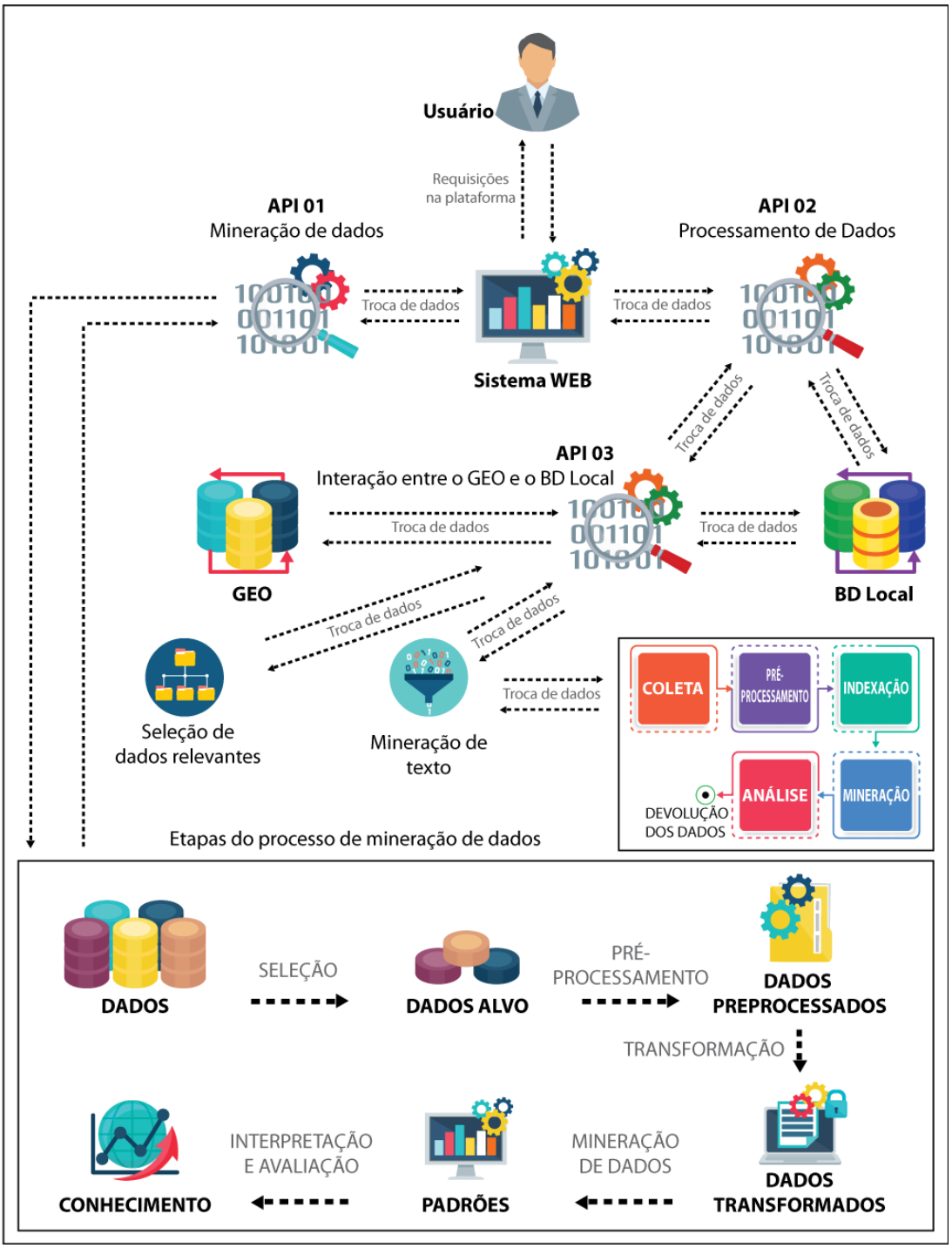
Com os dados devidamente filtrados e os resultados devidamente expostos ao pesquisador, o sistema disponibiliza uma serie de filtros voltados para a seleção de dados úteis para tentar identificar informações relevantes, objetivando novas descobertas científicas. Entre as métricas que podem ser utilizadas na análise dos resultados está o *Fold Change*, essa métrica avalia o quanto determinado gene foi expresso a mais ou a menos comparando amostras de dois indivíduos ou dois grupos de pacientes dentro de uma determinada pesquisa.

3.2 Arquitetura do sistema Mining_RNA

Conforme demonstrado na Figura 9, a arquitetura do sistema possui 3 API's distintas, um banco de dados e uma interface WEB na qual os usuários podem fazer consultas aos dados disponibilizados pelo sistema. Essa divisão em API's e interfaces foi motivada pela possibilidade de alocar os algoritmos em servidores diferentes caso seja necessário.

Alguns algoritmos que fazem parte das API's podem ser demasiadamente complexos do ponto de vista computacional, e isso pode acarretar em uma carga de processamento muito alta. Para que tudo seja melhor escalonado a arquitetura descentralizada se apresentou como uma solução eficiente.

Figura 9 – Visão Geral da Arquitetura do Sistema Mining_RNA



Fonte: Autoria Própria

A API 01 é responsável pelas tarefas relacionada a mineração dos dados obtidos, bem como das tarefas de aprendizagem de máquina. Os algoritmos implementados nessa

API possibilitam que os dados sejam validados antes do retorno ao usuário. Para essa validação interna, foi utilizado o algoritmo *Random Forest*. O uso deste se justifica pela sua adequação quando o número de atributos é maior que o número de amostras em um conjunto de dados, essa é uma forte característica dos microarranjos de DNA. É característica dessa API também que ela suporte a implementação de outros algoritmos de mineração em trabalhos futuros que venham a estender essa pesquisa.

A API 02 é responsável por pré-processar os dados que são carregados no sistema WEB. A API em questão pode ainda utilizar diretamente dados do banco de dados local, bem como requisitar que a API 03 busque novos dados no banco de dados biológicos GEO. O desenvolvimento dessa API foi feito respeitando rigorosamente princípios relacionados ao reuso para que assim, novos filtros possam ser inseridos em implementações futuras do projeto.

A API 03 realiza a captura dos dados do *dataset* solicitado pelo usuário e armazena as informações relevantes do mesmo no banco de dados local. Embora seja uma tarefa simples, em algumas ocasiões é necessário um poder de processamento considerável, pois os dados de algumas pesquisas são demasiadamente complexos. A API interpreta todos esses dados e os armazena para que assim as etapas posteriores de processamento fiquem mais rápidas. As três API's previamente descritas foram desenvolvidas utilizando a linguagem de programação Python, possibilitando com isso um ganho de desempenho quando comparada com linguagens web mais tradicionais como o PHP.

O sistema WEB é a interface que recebe todas as requisições do usuário. Este artefato foi desenvolvido em PHP para que sua disponibilização online pudesse ser facilitada. Os procedimentos a serem realizados nela são moderadamente facilitados, dessa forma filtros importantes não foram deixados de lado por serem complexos, é uma interface intuitiva e com mecanismos de ajuda para o usuário. Pretende-se que seja um sistema adaptável, mas no momento não faz parte do escopo uma interface específica para dispositivos com resolução pequena. Essa escolha se justifica pela densidade dos dados a serem exibidos, necessitando assim de uma resolução maior para uma visualização inteligível.

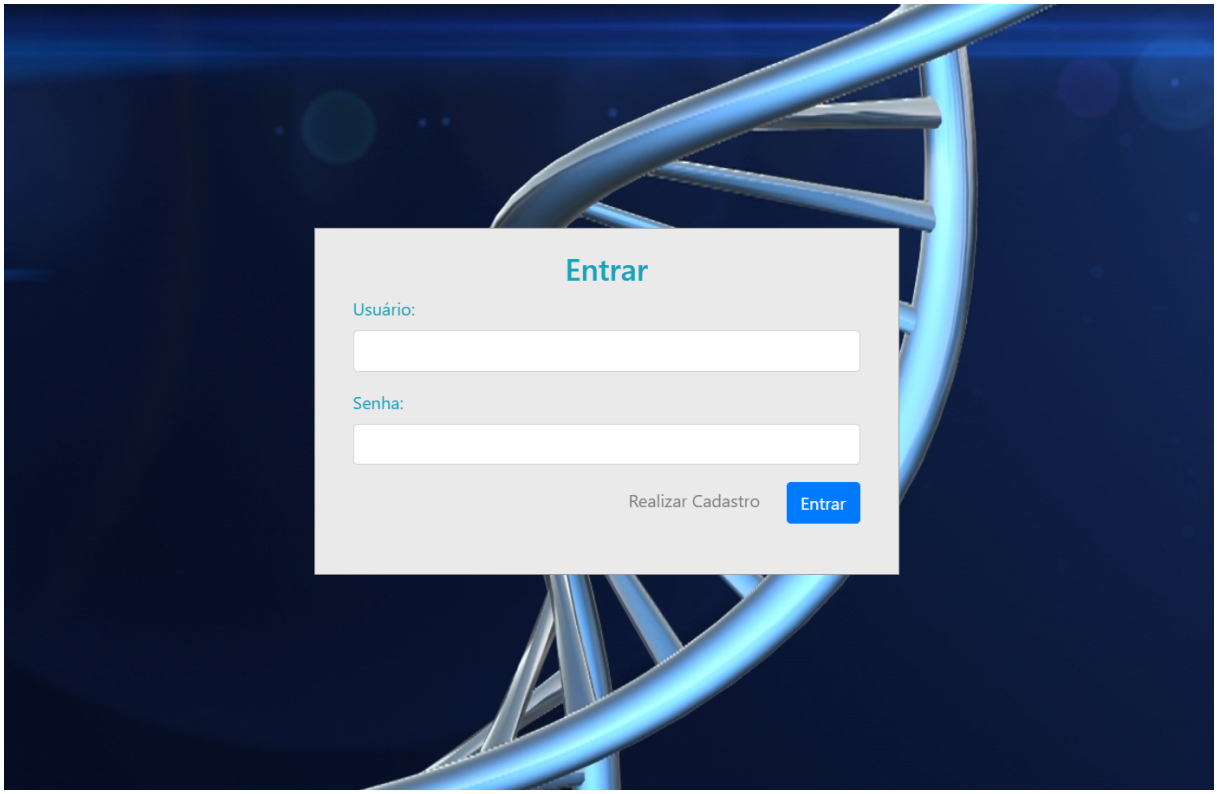
Os usuários do sistema precisam estar devidamente cadastrados para que dessa forma, possa haver uma personalização adequada, bem como possibilitar análises relevantes para a tomada de decisão relacionada a alocação de recursos e, implementações futuras. Essa autenticação pretende também prevenir abusos por parte do usuário, para assim preservar a disponibilidade e integridade do sistema desenvolvido.

Por fim, a base de dados relacional utilizada no sistema armazena todos os dados que são relevantes para as consultas. Essa base permite que alguns dados processados fiquem armazenados de forma que possibilite ao usuário recuperá-los em um momento futuro, armazenando também *presets* de filtros definidos pelo usuário para que ele possa aplicar novamente os mesmos filtros em situações futuras.

3.3 Utilização do sistema Mining_RNA

Esta seção descreve o sistema WEB desenvolvido durante esta pesquisa, denominado Mining_RNA. O sistema busca por meio de uma interface gráfica intuitiva fornecer dados relevantes a partir do estudo selecionado pelo pesquisador. Inicialmente o usuário precisa estar devidamente autenticado, conforme a Figura 10.

Figura 10 – Tela de Autenticação do sistema Mining_RNA



Fonte: Autoria Própria

Após a devida autenticação, o pesquisador deverá informar ao sistema qual o código da pesquisa que ele deseja analisar, assim como na Figura 11, os estudos armazenados no GEO podem ser acessados por meio dos códigos GSE ou GDS seguidos do identificador da pesquisa. Em ambos os casos, os dados são semelhantes, sendo que os que começam com GDS já passaram por um processo de curadoria por parte do NCBI. Embora seja possível a realização da comparação entre mais de um estudo simultaneamente, esta pesquisa está focada na análise de um *dataset* por vez. Ainda conforme a Figura 11, o estudo escolhido durante essa análise foi o GDS3875, uma pesquisa sobre diabetes tipo 1 e 2 que faz parte da curadoria do GEO.

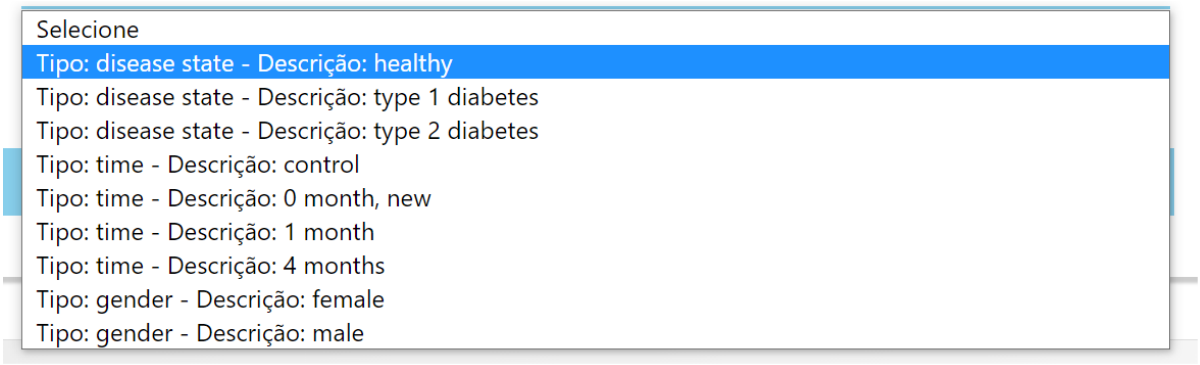
Figura 11 – Tela de seleção de estudo do sistema Mining_RNA

Fonte: Autoria Própria

Após a seleção do conjunto de dados o sistema fará uso da API 02 para acessar as informações solicitadas, caso a pesquisa escolhida ainda não faça parte do banco de dados local do sistema essa API solicitará que a API 03 acesse o banco biológico GEO e capture o arquivo desejado, armazenando-o localmente para que utilizações posteriores possam ser facilitadas.

Os estudos armazenados no GEO costumam dispor seus dados separados pelos grupos de pacientes que fizeram parte da pesquisa, sendo que na pesquisa selecionada os dados foram separados conforme o recorte apresentado na Figura 12.

Figura 12 – Grupos do estudo GDS3875

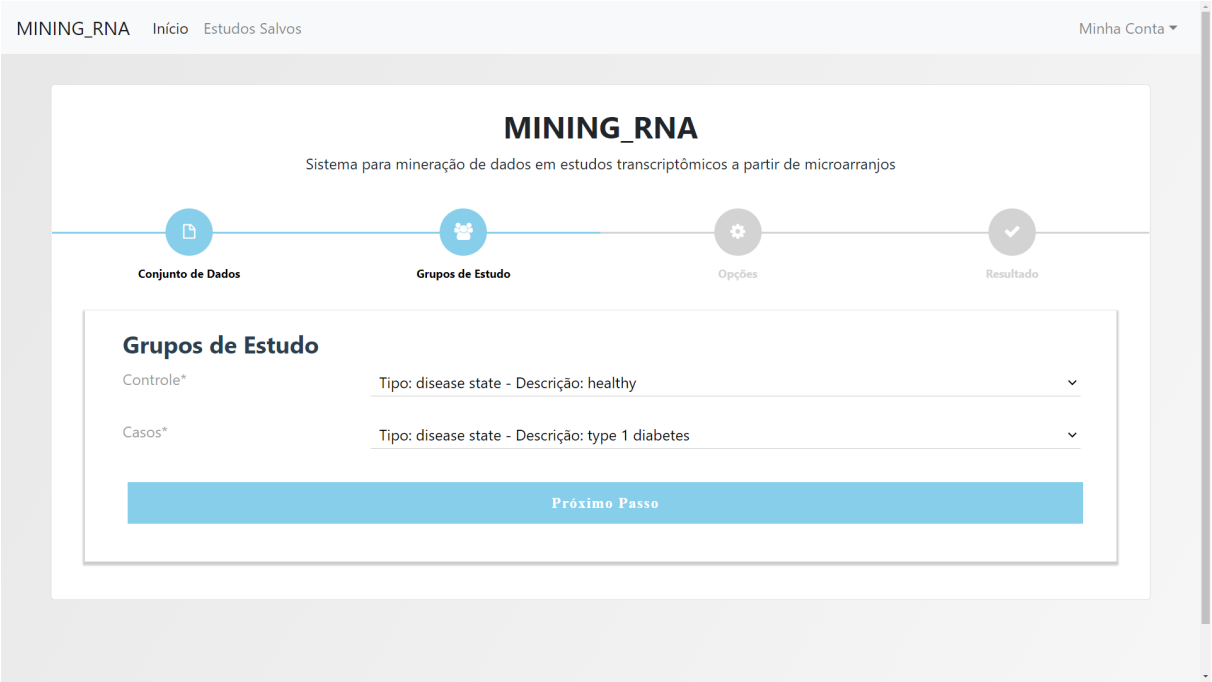


Fonte: Autoria Própria

Após a etapa inicial de pré-processamento o sistema disponibilizará os grupos da pesquisa de forma que o usuário do sistema poderá selecionar qual grupo usará como

controle e qual grupo contém os casos que ele deseja analisar, essa seleção é demonstrada na Figura 13 onde foi selecionado o subgrupo descrito como saudável como grupo de controle e o subgrupo descrito como portadores de diabetes tipo 1 como grupo a ser comparado (casos).

Figura 13 – Tela de seleção de grupo controle e grupo com casos do sistema Mining_RNA



Fonte: Autoria Própria

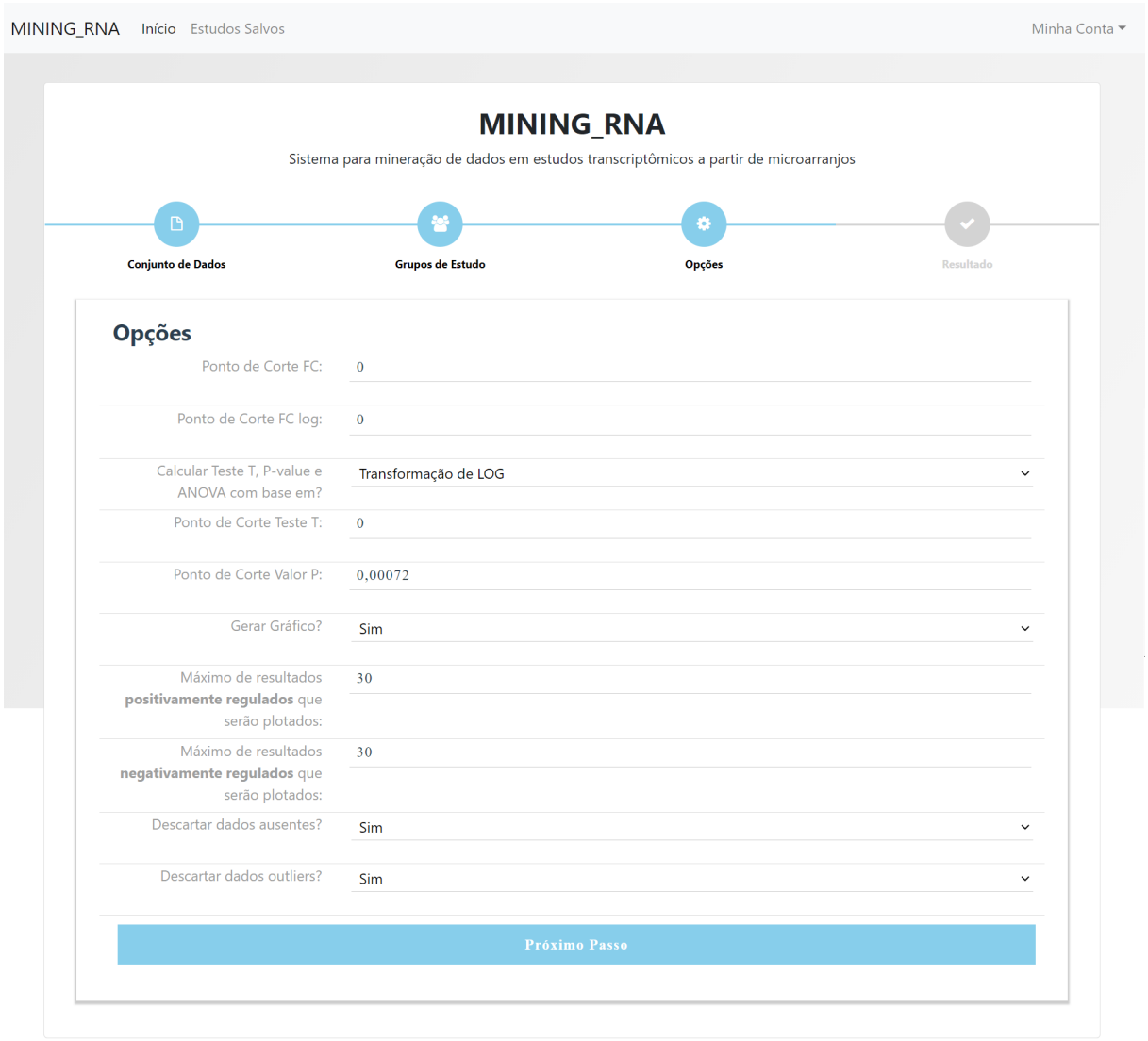
Depois dos grupos devidamente selecionados, o próximo passo, demonstrado na Figura 14 possibilita que o pesquisador personalize diversos parâmetros que filtrarão os dados do estudo para que o resultado traga informações relevantes. A opção "Ponto de corte FC" é responsável por limitar a exibição dos dados de *fold change* o valor selecionado será aplicado para os resultados de genes regulados acima (positivos) e regulados abaixo (negativos) na comparação das amostras do grupo de controle em relação as amostras do grupo de casos. O cálculo de *fold change* demonstra quanto determinado gene foi expresso em relação a outro. Também é possível definir um ponto de corte para o *fold change* que foi calculado utilizando a base logarítmica 2 através da opção "Ponto de corte FC Log".

Para possibilitar também uma análise estatística dos dados, o sistema faz cálculos utilizando Testes T e de Valor P. Este cálculo pode ser feito a partir dos dados brutos ou transformados em log_2. O valor do teste t é um dado importante já que cada grupo é composto por várias amostras, por isso, foi necessário que a média dos valores de cada gene dessas amostras fosse calculada, como médias estão sendo comparadas, este teste é importante para a confiabilidade dos dados. O cálculo do valor P é utilizado em diversos cálculos estatísticos, inclusive nos Testes T, para facilitar verificações feitas pelo

pesquisador o sistema já realiza esse cálculo. Ambos estes valores podem ter sua exibição limitada através das opções "Ponto de corte Teste T" e "Ponto de Corte Valor P".

É possível também que o sistema retorne um gráfico relacionado ao cálculo de *fold change*, essa opção também pode ser definida através dos controles expostos na Figura 14. Nesse gráfico o pesquisador pode definir quantos resultados serão exibidos para genes positivamente regulados e negativamente regulados. Ainda nessa tela o usuário pode escolher descartar dados ausentes e *outliers* que são dados significativamente fora da curva, essas opções são relevantes ao algoritmo de mineração de dados que é usado internamente na verificação dos resultados.

Figura 14 – Tela de filtros do sistema Mining_RNA

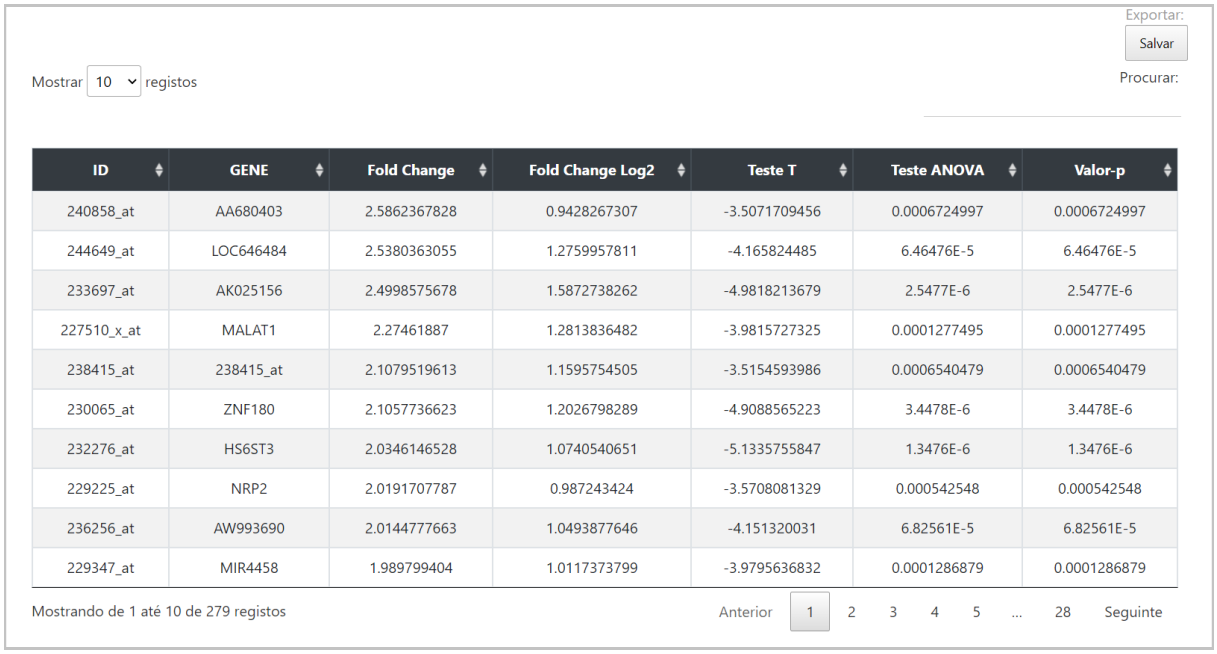


Fonte: Autoria Própria

Após a devida seleção dos filtros que devem aplicados, o sistema disponibiliza dois tipos de visualização dos dados calculados, a primeira é por meio de uma tabela conforme demonstra a Figura 15. Assim como pode ser visto, a tabela exibe respectivamente o

código identificador do gene, o nome do gene, o valor do *fold change* desse gene calculado entre os grupos escolhidos, o cálculo de *fold change* em log de 2, o valor do teste T, o valor do teste ANOVA e o valor-p. Inicialmente os dados são ordenados pelo valor do *fold change*, porém essa tabela é interativa e possibilita ao usuário ordenar os dados por qualquer uma das colunas disponibilizadas, é possível ainda aumentar a quantidade de registros exibidos em cada página da tabela e realizar uma pesquisa em qualquer dado disposto na tabela.

Figura 15 – Tela de resultados - Parte 1 (Tabela)

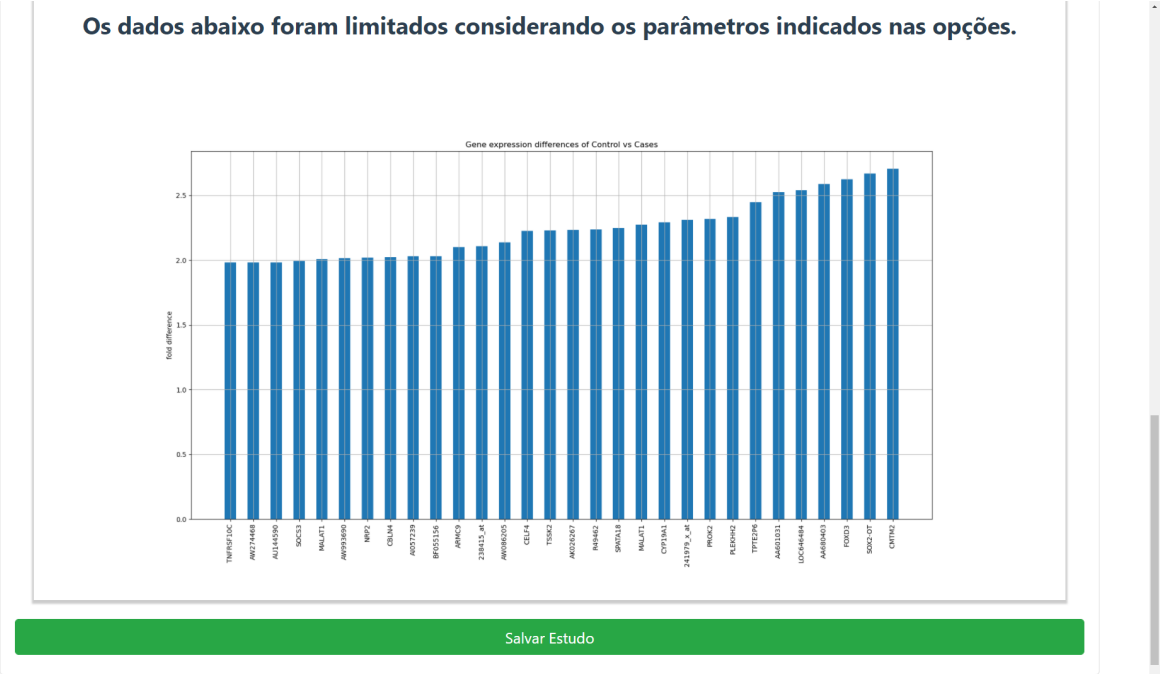


Fonte: Autoria Própria

É importante que o sistema possibilite que o pesquisador reutilize os dados obtidos em outras análises, então é disponibilizado juntamente a tabela de resultados, um botão que possibilita que os dados ali dispostos sejam exportados para um arquivo compatível com um editor de tabelas convencional, como o Microsoft Excel ou o OpenOffice Calc, ao selecionar essa opção os dados da tabela são imediatamente baixados para o computador do usuário. De posse desses dados o pesquisador pode adaptar o formado das informações para compatibiliza-las com outras aplicações de análise genética e assim aprofundar ainda mais sua pesquisa.

A segunda forma disponibilizada pelo sistema para a análise específica dos dados de *fold Change* é por meio do gráfico conforme Figura 16, esse gráfico é disponibilizado ao usuário na forma de uma imagem que além de disponível na tela, pode ser salva pelo pesquisador seguindo os passos padrões para esta ação, disponibilizados pelo próprio navegador no qual ele está acessando o sistema. Esse tipo de representação dos dados pode vir a ajudar no entendimento das informações ali dispostas ou ainda auxiliando que o mesmo possa envia-lo a terceiros ou utiliza-lo em apresentações dos resultados obtidos.

Figura 16 – Tela de resultados - Parte 2 (Gráfico)



Fonte: Autoria Própria

Ainda na tela de resultados é disponibilizado um botão para que o estudo em análise seja salvo na conta do usuário para que em um momento futuro esse estudo possa ser revisitado. Dessa forma o pesquisador não precisa repetir todos os passos ou anotar fisicamente os filtros escolhidos, com isso pretende-se agilizar o trabalho do usuário.

4 Validação

Este capítulo apresenta como foi realizada a validação do Mining_RNA e os resultados encontrados a partir da mesma. Para a execução dos testes foi selecionado um estudo a partir da base de dados biológicos GEO, este, está identificado pelo código GSE9006. A pesquisa em questão analisou a expressão gênica a partir de células mononucleares em crianças com diabetes (KAIZER *et al.*, 2007). O objetivo desta validação é comparar os resultados originalmente obtidos por Kaizer *et al.* (2007) com os resultados apresentados no Mining_RNA, para que isso fosse possível foram realizadas comparações utilizando o próprio artigo publicado a partir dos dados do GSE9006, mas também foram realizadas análises utilizando uma outra ferramenta disponibilizada pelo próprio NCBI, o GEO2R.

4.1 Comparação entre os resultados obtidos no Mining_RNA e no GEO2R utilizando o estudo GSE9006

Para os testes e validação da ferramenta objeto deste trabalho, foram carregadas as informações disponibilizadas pelo GSE9006 no Mining_RNA e no GEO2R, o estudo em questão disponibiliza no GEO duas plataformas com os dados brutos obtidos, o *Affymetrix Human Genome U133A Array* (código: GPL96) e o *Affymetrix Human Genome U133B Array* (código: GPL97) para este teste foram selecionados os resultados disponibilizados através do GPL96 nos dois sistemas utilizados. Essa seleção inicial é essencial pois os dados disponibilizados são diferentes entre as plataformas.

O estudo em questão analisou células mononucleares do sangue periférico de 43 pacientes com diabetes tipo 1 recém-diagnosticado, 12 pacientes com diabetes tipo 2 recém-diagnosticado e 24 controles saudáveis. Foram realizados ainda estudos através de amostras de acompanhamento de um e 4 meses obtidas a partir de 20 dos pacientes com diabetes tipo 1 (KAIZER *et al.*, 2007). Para esta validação foram selecionados o grupo de controle e o grupo com diabetes tipo 1 recém-diagnosticado o qual denominamos de "casos". Esses dois grupos serão comparados com o objetivo de identificar o *fold change* de cada gene expresso nas amostras dos dois grupos.

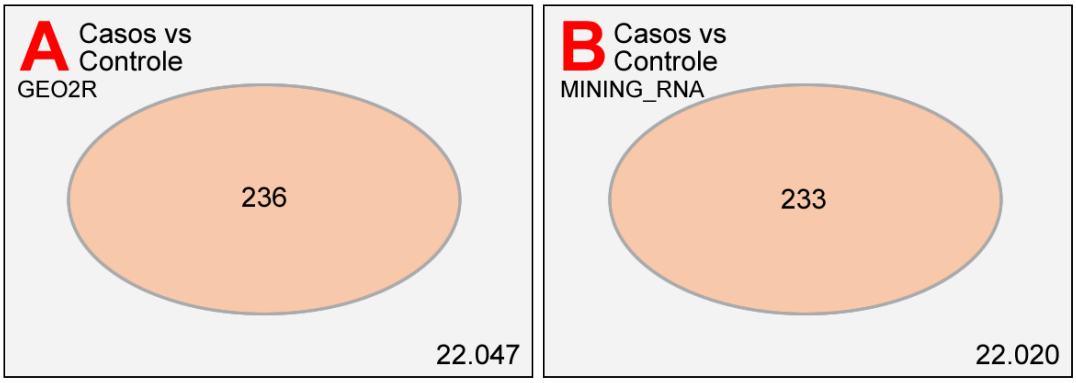
Para limitar a pesquisa à genes mais significativos foi definido um ponto de corte para o valor-P dos resultados obtidos em cada gene, esse ponto de corte foi estabelecido como sendo 0,00072 pois é um dos pontos de corte utilizados na pesquisa original. A partir desse fator limitador foram realizados testes utilizando os dados brutos obtidos através de microarranjo e esses mesmos valores após a aplicação da transformação para log_2, a análise com os dados transformados tem bastante relevância já que bibliotecas consolidadas em

R para análises biológicas como o limma[1], necessitam desses valores dessa forma para realizar seus cálculos de *fold change*.

4.1.1 Resultados obtidos através da análise dos dados brutos

Ao iniciar a análise foi percebido que havia uma pequena diferença entre a quantidade de genes apresentados pelo sistema GEO2R e Mining_RNA, no entanto essa diferença foi de somente aproximadamente 1,2%, sendo assim considerada uma diferença válida considerando que os algoritmos utilizados e linguagens de programação utilizados pelos sistemas são diferentes. O diagrama de Venn demonstrado através da Figura 17A representa o subconjunto de dados filtrado através no GEO2R e o diagrama apresentado na Figura 17B o subconjunto obtido através do Mining_RNA. Na figura em questão é possível perceber que foram selecionados respectivamente 236 e 233 genes que atendem aos parâmetros de corte definidos para este teste, sendo descartados respectivamente 22.047 e 22.020 genes nas análises dos sistemas GEO2R e Mining_RNA.

Figura 17 – Diagrama de Venn dos genes selecionados através dos dados brutos



Fonte: Autoria Própria

A Tabela 3 apresenta esses resultados separando os genes obtidos em cada sistema em regulados acima e abaixo na comparação entre o grupo de controle e o de casos. Nessa tabela é possível perceber que há uma concentração maior de genes regulados abaixo, ou seja, são genes que são superexpressos nas amostras do grupo de casos, tendo uma expressão menor no grupo de controle.

Tabela 3 – Quantidade de genes regulados acima e abaixo comparando os dois sistemas

	Regulados acima	Regulados abaixo
Sistema GEO2R	79	157
Sistema Mining_RNA	78	155

Fonte: Autoria Própria

[1] https://bioconductor.org/packages/release/bioc/html/limma.html

Para melhor validar os resultados, a Tabela 4 apresenta a matriz de comparação para os genes regulados acima e a Tabela 5 apresenta a matriz para os que foram regulados abaixo considerando os grupos de controle e de casos selecionados na aplicação. A identificação dos valores para esses genes foi realizada através do cálculo de *fold change* feito a partir dos dados brutos.

Tabela 4 – Matriz de comparação entre os genes regulados acima nas duas plataformas

Situação	**Quantidade**
Regulados acima nas duas plataformas	78
Regulado acima no GEO2R e não regulado acima no Mining_RNA	1
Regulado acima no Mining_RNA e não regulado acima no GEO2R	0

Fonte: Autoria Própria

Tabela 5 – Matriz de comparação entre os genes regulados abaixo nas duas plataformas

Situação	**Quantidade**
Regulados abaixo nas duas plataformas	153
Regulado abaixo no GEO2R e não regulado acima no Mining_RNA	4
Regulado abaixo no Mining_RNA e não regulado acima no GEO2R	2

Fonte: Autoria Própria

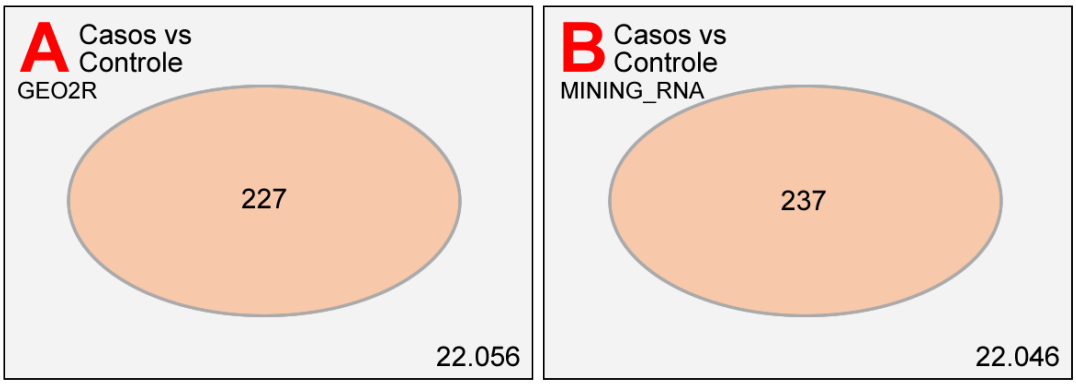
A partir dos valores apresentados pelas duas plataformas é possível identificar que a diferença entre os resultados dos dois sistemas pequena. Mesmo com os dados sem uma normalização relevante os algoritmos das duas plataformas conseguiram uma eficiência maior que 97% considerando os valores positivos, falsos-positivos e falsos-negativos dos genes regulados acima e abaixo.

4.1.2 Resultados obtidos através da análise dos dados ajustados para $\log_2$

Outro teste relevante foi realizado a partir dos dados ajustados para log_2, a transformação de log torna dados mais simétricos, diminuindo a distorção em comparação com os níveis de expressão gênica em escala linear e, portanto, um teste estatístico paramétrico fornecerá uma resposta mais precisa e relevante.

Na análise dos dados ajustados para log_2 houve um sensível aumento entre a quantidade de genes apresentada entre os dois sistemas, nesse teste essa diferença foi de 4,4%. O diagrama de Venn demonstrado através da Figura 18A representa o subconjunto de dados filtrado através no GEO2R e o diagrama apresentado na Figura 18B o subconjunto obtido através do Mining_RNA. Nos respectivos diagramas é possível perceber que foram selecionados 227 e 237 genes que atendem aos parâmetros de corte definidos para este teste, sendo descartados respectivamente 22.056 e 22.046 genes nas análises dos sistemas GEO2R e Mining_RNA.

Figura 18 – Diagrama de Venn dos genes selecionados através dos dados brutos



Fonte: Autoria Própria

A Tabela 6 apresenta esses resultados separando os genes obtidos em cada sistema em regulados acima e abaixo na comparação entre o grupo de controle e o de casos. Nessa tabela é possível perceber que houve um maior equilíbrio entre a regulação genica, isso é possível considerando a transformação de log aplicada, a qual diminuiu a distorção entre as expressões analisadas. Os resultados demonstrados a seguir foram realizados considerando os valores apresentados no cálculo de log de *fold change*.

Tabela 6 – Quantidade de genes regulados acima e abaixo comparando os dois sistemas

	Regulados acima	Regulados abaixo
Sistema GEO2R	128	99
Sistema Mining_RNA	132	105

Fonte: Autoria Própria

Com o objetivo de entender melhor o deslocamento entre os valores de expressão genica, a Tabela 7 apresenta a matriz de comparação para os genes regulados acima e a Tabela 8 apresenta a matriz para os que foram regulados abaixo considerando os grupos de controle e de casos selecionados na aplicação.

Tabela 7 – Matriz de comparação entre os genes regulados acima nas duas plataformas

Situação	**Quantidade**
Regulados acima nas duas plataformas	125
Regulado acima no GEO2R e não regulado acima no Mining_RNA	3
Regulado acima no Mining_RNA e não regulado acima no GEO2R	7

Fonte: Autoria Própria

Tabela 8 – Matriz de comparação entre os genes regulados abaixo nas duas plataformas

Situação	**Quantidade**
Regulados abaixo nas duas plataformas	96
Regulado abaixo no GEO2R e não regulado acima no Mining_RNA	3
Regulado abaixo no Mining_RNA e não regulado acima no GEO2R	9

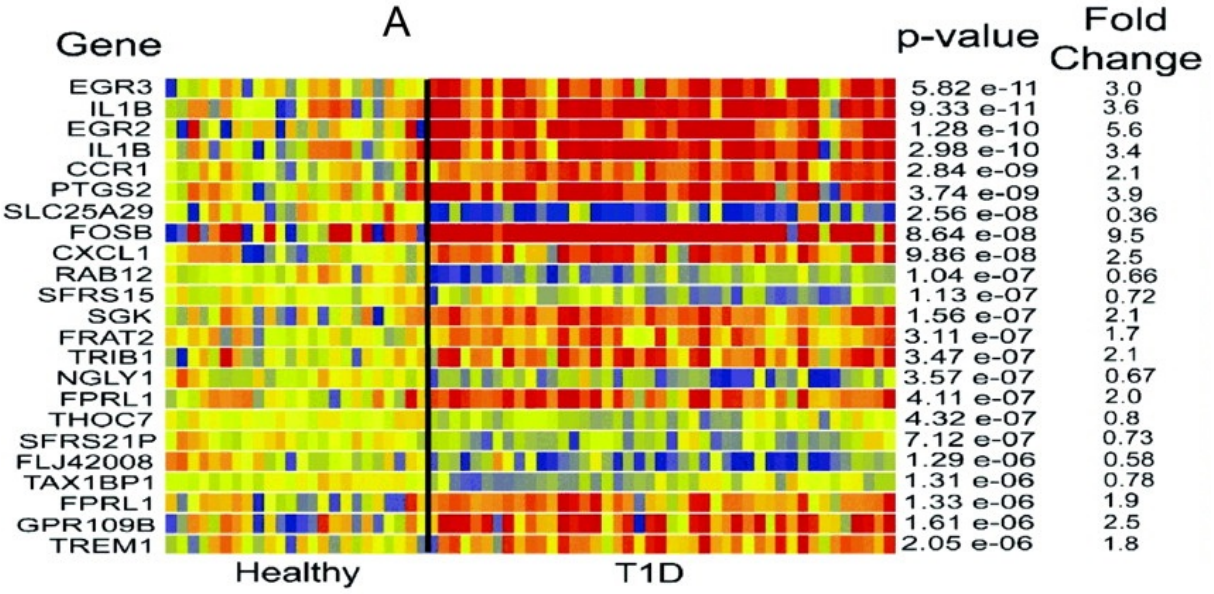
Fonte: Autoria Própria

Considerando os valores apresentados pelas duas plataformas é possível identificar que a diferença entre os resultados dos dois sistemas ainda é pequena se comparada com os testes realizados utilizando os dados brutos. Quando aplicada a transformação logarítmica nos dados do estudo os algoritmos das duas plataformas conseguiram uma eficiência maior que 90% considerando os valores positivos, falsos-positivos e falsos-negativos dos genes regulados acima e abaixo.

4.2 Comparação entre resultados de expressão diferencial entre o estudo original e os resultados do Mining_RNA

O terceiro teste feito para validar o objeto deste estudo foi comparar um dos resultados obtidos na pesquisa original com os dados obtidos através do Mining_RNA. A Figura 19 apresenta um subgrupo de genes que foi obtido através de um ponto de corte bem restrito, este, buscou obter os genes mais significativos para o estudo em questão. O teste realizado no presente estudo buscou identificar a diferença na expressão gênica entre os genes apresentados na figura em comparação com os mesmos genes obtidos através do sistema Mining_RNA. Para este teste foram considerados tanto os dados armazenados na plataforma GPL96 quanto os da plataforma GPL97, isto foi necessário porque na análise original o pesquisador utilizou as duas plataformas.

Figura 19 – Mapa de calor disponibilizado no estudo original contendo o *fold change* para cada gene na comparação entre grupo de controle (Healthy) e casos (T1D)

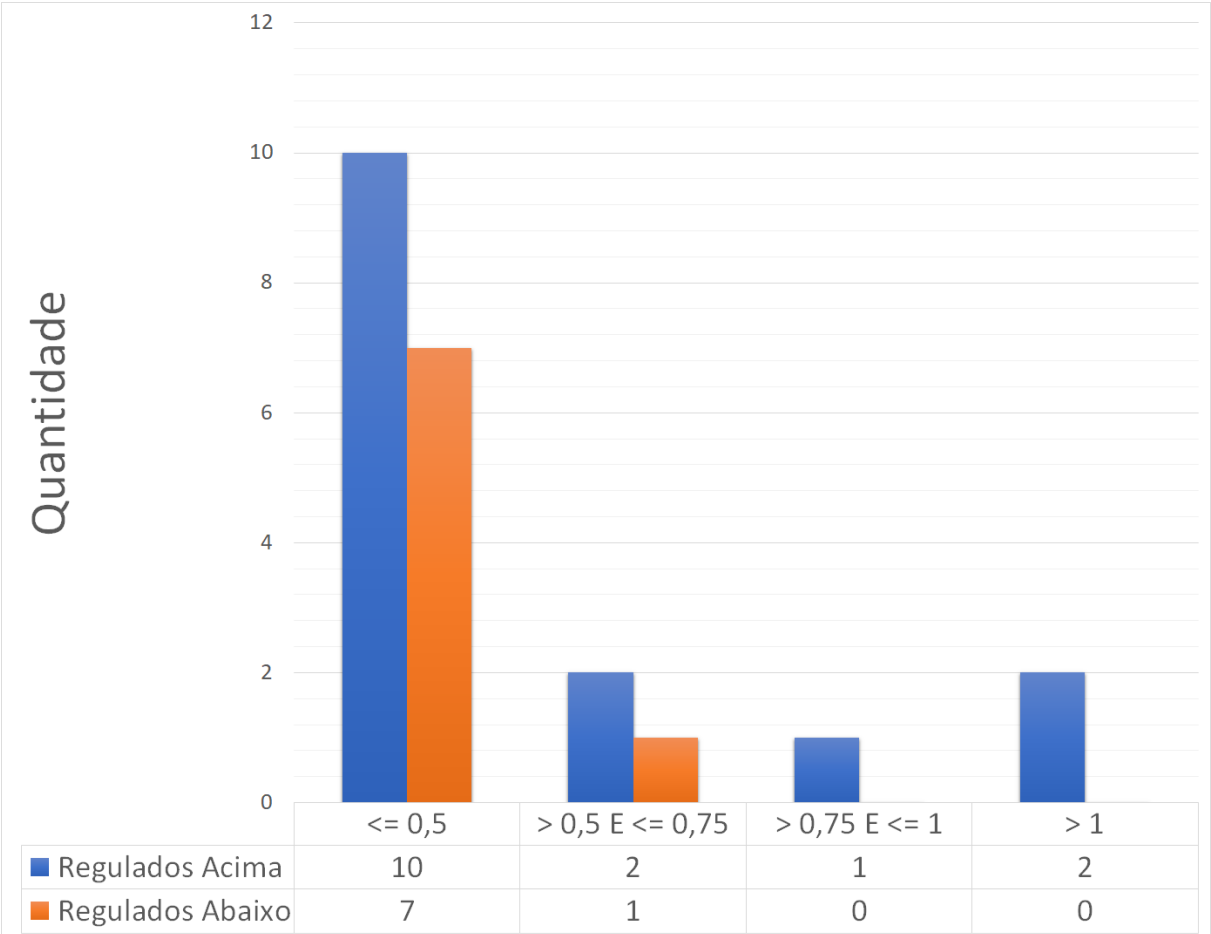


Fonte: Adaptado de Kaizer *et al.* (2007)

A Figura 20 demonstra um gráfico de barras com a avaliação da proximidade dos resultados obtidos pelo Mining_RNA com os resultados da pesquisa original (Figura 19.

Foram listados na figura em questão as diferenças obtidas tanto para os genes superexpressos quanto para os subexpressos. A partir da análise do gráfico é possível identificar que em aproximadamente 69,5% dos genes avaliados, a diferença é menor que 0,5 na comparação entre o resultado original e o calculado por este estudo. Em aproximadamente 13%, a diferença entre os *fold changes* ficaram entre 0,5 e 0,75. Somente 1 dos 23 genes avaliados teve a disparidade entre 0,75 e 1 (Aproximadamente 4%). Só foi possível identificar maiores diferenças na comparação de dois genes superexpressos, pouco mais de 8% dos genes avaliados, estes, tiveram suas diferenças calculadas em mais de 1 na comparação entre os dois estudos.

Figura 20 – Diferença entre valores de *fold change* obtidos pelo Mining_RNA e os encontrados no estudo original



Fonte: Autoria Própria

4.3 Conclusão da Validação

Após as análises realizadas a partir dos dados do estudo GSE9006, foi possível perceber que o nível de eficiência apresentado pelo Mining_RNA é satisfatório. A Tabela 9 sintetiza as eficiências obtidas nos três testes realizados para a validação do sistema.

Tabela 9 – Resumo de eficiência da validação

Teste Realizado	Eficiência
Eficiência calculada a partir dos resultados obtidos através da análise dos dados brutos a partir do GEO2R e o Mining_RNA.	99%
Eficiência calculada a partir dos resultados obtidos através da análise dos dados ajustados para log_2 a partir do GEO2R e o Mining_RNA.	96%
Eficiência calculada considerando as diferenças menores que 1, a partir das comparações entre os resultados de expressão diferencial do estudo original e os resultados do Mining_RNA.	91%

Fonte: Autoria Própria

No primeiro teste o sistema obteve uma eficiência aproximada de 99% na comparação com um sistema mantido pelo próprio NCBI. No segundo teste, utilizando o mesmo conjunto de dados com pré-processamento diferente a eficácia foi de aproximadamente 96% a qual ainda é alta considerando a utilização de algoritmos e linguagem de programação diferentes, sendo o sistema comparado com outro com maturidade significativamente maior. Os dois testes evidenciam uma eficiência média de 97,5% no cenário proposto.

O terceiro teste, realizado comparando os resultados calculados a partir do objeto produzido por esta pesquisa com o estudo original evidenciou que em mais de 91% dos genes avaliados a diferença no valor do *fold change* calculado foi menor que 1 como pode ser visto na Figura 20, este resultado possibilita uma validação satisfatória para os resultados obtidos por esta pesquisa, mostrando ainda que os as informações obtidas através do sistema são confiáveis pois são significativamente próximas às obtidas através de outros tipos de estudos já publicados em meios científicos.

5 Considerações Finais

Este trabalho apresentou o Mining_RNA, um sistema web para mineração de dados em estudos transcriptômicos a partir de microarranjos, com interface de fácil acesso. O sistema possibilita aos pesquisadores a realização de reanálises em estudos armazenados no banco de dados biológicos GEO.

Durante a etapa de planejamento desse trabalho foram levantadas questões à partir da metodologia *design science* que foram devidamente satisfeitas e documentadas no decorrer da execução do projeto. Foi construído um sistema que, através de alguns passos e a devida seleção de filtros, oportuniza a pesquisadores reanalisar resultados de aproximadamente 143 mil (NCBI, 2021b) estudos transcriptômicos sem que seja necessário entender de programação, promovendo ainda um ganho de tempo que pode ser dedicado a análise final dos dados apresentados pelo sistema.

Para que os resultados apresentados no sistema desenvolvido pudessem ser confirmados como satisfatórios, foi executada a validação dos mesmos comparando com valores gerados através da ferramenta GEO2R do NCBI e confrontando também com os resultados de um artigo com resultados devidamente publicados em meios científicos. Os resultados gerados pelo Mining_RNA mostraram-se satisfatórios e com nível de confiabilidade alta.

Como perspectivas futuras no desenvolvimento de novas versões sistema Mining_RNA, estudos podem vir a explorar novas funcionalidades do sistema:

- Implementar algoritmos de mineração de texto para uma seleção mais inteligente do grupo de controle e de casos;
- Possibilitar a geração de outros tipos de representação gráfica dos resultados;
- Viabilizar a análise de dados de microarranjos ainda não publicados;
- Propiciar a análise de dados de microarranjos de outros bancos de dados biológicos;
- Desenvolver uma versão do sistema para execução local;
- Implementar o cálculo de *false discovery rate*;
- Possibilitar a exportação dos dados para o sistema Enrichr.

Referências

ALEXANDER, I. F. A taxonomy of stakeholders: Human roles in system development. *IJTHI*, v. 1, p. 23–59, 01 2005. Citado na página 13.

ALVES, E. A.; SOUZA, D. S. *et al.* Biología molecular. In: . [S.l.]: EPSJV, 2013. Citado na página 12.

BANAGANAPALLI, B.; SHAIK, N. A. Introduction to bioinformatics. In: *Essentials of Bioinformatics, Volume I*. [S.l.]: Springer, 2019. p. 1–18. Citado 2 vezes nas páginas 19 e 20.

BAXEVANIS, A. D.; BADER, G. D.; WISHART, D. S. *Bioinformatics: a practical guide to the analysis of genes and proteins*. 4. ed. [S.l.]: Wiley, 2020. Citado na página 19.

BENNETT, P.; HARDIKER, N. R. The use of computerized clinical decision support systems in emergency care: a substantive review of the literature. *Journal of the American Medical Informatics Association*, v. 24, n. 3, p. 655–668, 12 2016. ISSN 1067-5027. Disponível em: <https://doi.org/10.1093/jamia/ocw151>. Citado na página 15.

BONO, H. All of gene expression (aoe): an integrated index for public gene expression databases. *PloS one*, Public Library of Science San Francisco, CA USA, v. 15, n. 1, p. e0227076, 2020. Citado na página 34.

BRAZMA, A. Minimum information about a microarray experiment (miame) – successes, failures, challenges. *TheScientificWorldJournal*, v. 9, p. 420–3, 02 2009. Citado 2 vezes nas páginas 13 e 18.

CHEN, G. *et al.* Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. *Database*, v. 2019, 01 2019. ISSN 1758-0463. Bay145. Disponível em: <https://doi.org/10.1093/database/bay145>. Citado 2 vezes nas páginas 27 e 28.

CHEN, Y.-L.; CHEN, J.-M.; TUNG, C.-W. A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. *Decision Support Systems*, v. 42, n. 3, p. 1503 – 1520, 2006. ISSN 0167-9236. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167923606000030>. Citado na página 24.

DAVIS, S.; MELTZER, P. S. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, Oxford University Press, v. 23, n. 14, p. 1846–1847, 2007. Citado 2 vezes nas páginas 27 e 29.

DJORDJEVIC, D. *et al.* Discovery of perturbation gene targets via free text metadata mining in gene expression omnibus. *Computational Biology and Chemistry*, v. 80, p. 152 – 158, 2019. ISSN 1476-9271. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1476927119301963>. Citado 2 vezes nas páginas 30 e 31.

DORFMAN, L. E. *et al.* Microarray-based comparative genomic hybridization analysis in neonates with congenital anomalies: detection of chromosomal imbalances. *Jornal de Pediatria*, scielo, v. 91, p. 59 – 67, 02 2015. ISSN 0021-7557. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0021-75572015000100059&nrm=iso>. Citado na página 21.

ESPINDOLA, F. S. *et al.* Recursos de bioinformática aplicados às ciências ômicas como genômica, transcriptômica, proteômica, interatômica e metabolômica. *Bioscience Journal*, v. 26, n. 3, 2010. Citado na página 19.

GANGWAR, V.; GHOSE, U.; SINGH, Y. Data mining of biological data in bioinformatics using transcription, translation algorithm and pattern matching of protein sequences. *International Journal of Advanced Research in Computer Science*, International Journal of Advanced Research in Computer Science, v. 3, n. 3, 2012. Citado na página 26.

GARG, S. B.; MAHAJAN, A. K.; KAMAL, T. An approach for diabetes detection using data mining classification techniques. *Journal of Engineering Sciences*, v. 26, 2017. ISSN 2320-0332. Citado na página 26.

GASPAROVICA-ASĪTE, M.; ALEKSEJEVA, L. Classification methodology for bioinformatics data analysis. *Automatic Control and Computer Sciences*, Springer, v. 53, n. 1, p. 28–38, 2019. Citado na página 15.

GELDER, C. W. V. *et al.* Bioinformatics in the netherlands: the value of a nationwide community. *Briefings in bioinformatics*, Oxford University Press, v. 20, n. 2, p. 375–383, 2019. Citado na página 19.

GONZALO, R.; SÁNCHEZ, A. Introduction to microarrays technology and data analysis. In: *Comprehensive Analytical Chemistry*. [S.l.]: Elsevier, 2018. v. 82, p. 37–69. Citado na página 20.

GOVINDARAJAN, R. *et al.* Microarray and its applications. *Journal of pharmacy & bioallied sciences*, Medknow Publications & Media Pvt Ltd, v. 4, n. Suppl 2, p. S310–S312, Aug 2012. ISSN 0975-7406. 23066278[pmid]. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/23066278>. Citado 2 vezes nas páginas 20 e 21.

GREENE, C. S.; TROYANSKAYA, O. G. PILGRM: an interactive data-driven discovery platform for expert biologists. *Nucleic Acids Research*, v. 39, p. W368–W374, 06 2011. ISSN 0305-1048. Disponível em: <https://doi.org/10.1093/nar/gkr440>. Citado na página 20.

GUDENAS, B. L. *et al.* Genomic data mining for functional annotation of human long noncoding rnas. *Journal of Zhejiang University-SCIENCE B*, v. 20, n. 6, p. 476–487, Jun 2019. ISSN 1862-1783. Disponível em: <https://doi.org/10.1631/jzus.B1900162>. Citado na página 15.

GUINDALINI, C.; TUFIK, S. Uso de microarrays na busca de perfis de expressão gênica - aplicação no estudo de fenótipos complexos. *Brazilian Journal of Psychiatry*, scielo, v. 29, p. 370 – 374, 12 2007. ISSN 1516-4446. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-44462007000400014&nrm=iso>. Citado na página 21.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790. Citado na página 25.

HERNÁNDEZ-CABRONERO, M. *et al.* Analysis-driven lossy compression of dna microarray images. *IEEE Transactions on Medical Imaging*, v. 35, n. 2, p. 654–664, Feb 2016. ISSN 0278-0062. Citado na página 12.

HIRA, Z.; GILLIES, D. F. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, v. 2015, p. 1–13, 07 2015. Citado 2 vezes nas páginas 12 e 26.

HOGEWEG, P. The roots of bioinformatics in theoretical biology. *PLoS computational biology*, Public Library of Science, v. 7, n. 3, p. e1002021–e1002021, Mar 2011. ISSN 1553-7358. 21483479[pmid]. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/21483479>. Citado na página 19.

HUERTA, M. *et al.* Mgdb: Crossing the marker genes of a user microarray with a database of public-microarrays marker genes. *Bioinformatics (Oxford, England)*, v. 30, 02 2014. Citado na página 13.

HUSSAIN, A.; CAMBRIA, E. Semi-supervised learning for big social data analysis. *Neurocomputing*, v. 275, p. 1662 – 1673, 2018. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231217316363>. Citado na página 25.

KAIZER, E. C. *et al.* Gene Expression in Peripheral Blood Mononuclear Cells from Children with Diabetes. *The Journal of Clinical Endocrinology Metabolism*, v. 92, n. 9, p. 3705–3711, 09 2007. ISSN 0021-972X. Disponível em: <https://doi.org/10.1210/jc.2007-0979>. Citado 2 vezes nas páginas 44 e 48.

KOEPPEN, K.; STANTON, B. A.; HAMPTON, T. H. ScanGEO: parallel mining of high-throughput gene expression data. *Bioinformatics*, v. 33, n. 21, p. 3500–3501, 07 2017. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btx452>. Citado na página 30.

LAN, K. *et al.* A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, Springer, v. 42, n. 8, p. 139, 2018. Citado na página 26.

MISSOUM, A. Dna microarray and bioinformatics technologies: A mini-review. *Proc. Nat. Res. Soc*, v. 2, p. 02010, 2018. Citado na página 20.

MITCHELL, T. M. *Machine Learning*. 1. ed. USA: McGraw-Hill, Inc., 1997. ISBN 0070428077. Citado na página 25.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. *About GEO2R*. 2021. Consultado em janeiro de 2021. Disponível em: <https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html>. Citado 2 vezes nas páginas 31 e 32.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. *GEO - NCBI*. 2021. Consultado em janeiro de 2021. Disponível em: <https://www.ncbi.nlm.nih.gov/geo/>. Citado na página 51.

NIE, H. *et al.* Microarray data mining using bioconductor packages. *BMC Proceedings*, Springer Verlag, v. 3, n. Suppl.4, p. S9, 2009. ISSN 1753-6561. Citado na página 26.

OZA, N. C. *Ensemble Data Mining Methods*. 2019. Data do acesso: 15 nov. 2019. Disponível em: <https://catalog.data.gov/dataset/ensemble-data-mining-methods>. Acesso em: 15 nov. 2019. Citado na página 25.

PAPATHEODOROU, I. *et al.* Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Research*, v. 46, n. D1, p. D246–D251, 11 2017. ISSN 0305-1048. Disponível em: <https://doi.org/10.1093/nar/gkx1158>. Citado na página 12.

PEREIRA, J. L. *ANÁLISE PREDITIVA EM SISTEMAS DE INFORMAÇÃO NO CONTEXTO DO BIG DATA*. Marília: UNIVEM, 2014. Citado na página 23.

PIATESKI, G.; FRAWLEY, W. *Knowledge Discovery in Databases*. Cambridge, MA, USA: MIT Press, 1991. ISBN 0262660709. Citado na página 22.

PIERCE, B. A. *Genetics: A Conceptual Approach*. W. H. Freeman, 2016. ISBN 1319050964. Disponível em: <https://www.xarg.org/ref/a/1319050964/>. Citado na página 12.

PIETKA, E. *et al. Information Technology in Biomedicine*. [S.l.]: Springer, 2019. v. 1011. Citado na página 15.

ROSA, G. J. d. M.; ROCHA, L. B. d.; FURLAN, L. R. Estudos de expressão gênica utilizando-se microarrays: delineamento, análise, e aplicações na pesquisa zootécnica. *Revista Brasileira de Zootecnia*, scielo, v. 36, p. 186 – 209, 07 2007. ISSN 1516-3598. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-35982007001000018&nrm=iso>. Citado 2 vezes nas páginas 20 e 21.

RUNG, J.; BRAZMA, A. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, Nature Publishing Group, v. 14, n. 2, p. 89–99, 2013. Citado na página 26.

SCHMITZ-ABE, K. *et al.* Unique bioinformatic approach and comprehensive reanalysis improve diagnostic yield of clinical exomes. *European Journal of Human Genetics*, Nature Publishing Group, v. 27, n. 9, p. 1398–1405, 2019. Citado na página 13.

SEDKAOUI, S. *Data analytics and big data*. [S.l.]: John Wiley & Sons, 2018. Citado na página 22.

Seni, G.; Elder, J. [S.l.: s.n.], 2010. Citado na página 25.

SHMUELI, G. *et al. Data Mining for Business Analytics: Concepts, Techniques and Applications in Python*. [S.l.]: John Wiley & Sons, 2019. Citado na página 22.

SUN, M.-a.; SHAO, X.; WANG, Y. Microarray data analysis for transcriptome profiling. In: _____. *Transcriptome Data Analysis: Methods and Protocols*. New York, NY: Springer New York, 2018. p. 17–33. ISBN 978-1-4939-7710-9. Disponível em: <https://doi.org/10.1007/978-1-4939-7710-9_2>. Citado na página 12.

TORO-DOMíNGUEZ, D. *et al.* ImaGEO: integrative gene expression meta-analysis from GEO database. *Bioinformatics*, v. 35, n. 5, p. 880–882, 08 2018. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/bty721>. Citado 2 vezes nas páginas 28 e 29.

WANG, Z.; LACHMANN, A.; MA'AYAN, A. Mining data and metadata from the gene expression omnibus. *Biophysical Reviews*, v. 11, n. 1, p. 103–110, Feb 2019. ISSN 1867-2469. Disponível em: <https://doi.org/10.1007/s12551-018-0490-8>. Citado 2 vezes nas páginas 14 e 34.

WIERINGA, R. J. *Design Science Methodology for Information Systems and Software Engineering*. [S.l.: s.n.], 2014. 1-332 p. Citado 3 vezes nas páginas 13, 14 e 18.

WONG, K.-C. Big data challenges in genome informatics. *Biophysical Reviews*, v. 11, n. 1, p. 51–54, 2019. ISSN 1867-2469. Disponível em: <https://doi.org/10.1007/s12551-018-0493-5>. Citado na página 12.

WORKU, T.; NEGASSU, D. Review on dna micro array technologyand its application. *American Journal of Zoology*, v. 2, n. 4, p. 44–50, 2019. Citado na página 20.