UPF EDITORA UNIVERSIDADE DE PASSO FUNDO

RBCA ISSN 2176-6649

Revista Brasileira de Computação Aplicada, Novembro, 2021

DOI: 10.5335/rbca.v13i3.12285
Vol. 13, Nº 3, pp. 101–108
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Desenvolvimento e validação do sistema Mining_RNA

Development and validation of the Mining_RNA system

Carlos Renan Moreira [1], Christina Pacheco [2], Marcos Vinícius Pereira Diógenes [1], Cicília Raquel Maia Leite [1]

[1]UERN, Rua Almino Afonso, 478 - Centro - Mossoró, RN - Brasil - CEP: 59.610-210, [2]UnB, Brasília, DF - Brasil - CEP: 70910-900

*c.renan.moreira@gmail.com; christinaosvaldo@yahoo.com.br; marcosdiogenes@alu.uern.br; ciciliamaia@gmail.com

Recebido: 19/02/2021. **Revisado:** 27/10/2021. **Aceito:** 30/11/2021.

Resumo

O sequenciamento do genoma humano proporcionou o aprofundamento de diversos tipos de estudos e tecnologias de análise biológica, dentre estas, o microarranjo. A necessidade publicar os dados brutos dessas pesquisas impulsionou a criação de bancos de dados públicos onde essas informações pudessem ser indexadas e resgatadas. Essas bases são uma grande fonte de dados transcriptômicos que infelizmente acabam sendo subutilizadas. O objetivo deste trabalho foi o desenvolvimento de um sistema WEB para mineração de dados em estudos transcriptômicos a partir de microarranjos armazenados no banco de dados biológico Gene Expression Omnibus (GEO), o Mining_RNA. Através de uma usabilidade passo-a-passo juntamente com uma série de filtros o sistema possibilita resgatar dados do GEO, calcular a expressão diferencial entre os genes de um estudo, possibilitando ainda análises estatísticas para cada gene do estudo analisado. O sistema foi validado através da comparação com a avaliação dos mesmos dados com o software GEO2R (eficácia aproximada de 98%) e no estudo original (eficácia maior que 90%). Mining_RNA pode ser um forte aliado dos pesquisadores para a reanálise de estudos transcriptômicos, possibilitando uma nova forma de analisar os dados e gerando resultados tão confiáveis quanto ferramentas já consolidadas.

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

Abstract

The human genome sequencing provided the deepening of several types of studies and technologies of biological analysis, among these, the microarray. 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. Our aim was the development of a WEB system for data mining in transcriptomic studies from microarrays stored in the Gene Expression Omnibus biological database (GEO). Through a step-by-step usability together with a series of filters the Mining_RNA system makes it possible to retrieve GEO data, calculate the differential expression between the genes of a study, and statistical analysis for each gene of the analyzed study. The system was validated by comparing with the evaluation of the same data with the GEO2R software (approximately 98% effectiveness) and in the original study (effectiveness higher than 90%). Mining_RNA can be a strong ally for researchers for the re-analysis of transcriptomic studies, enabling a new way of analyzing data and generating results as reliable as already consolidated tools.

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

1 Introdução

A computação aplicada a estudos biológicos, principalmente os que focam no DNA (ácido desoxirribonucléico), RNA (ácido ribonucléico) e proteínas, vem contribuindo para o a aceleração e o alto rendimento das pesquisas genéticas, atuando no processamento de grandes quantidades de informações biológicas (Gasparovica-Asīte and Aleksejeva, 2019). Estudos de bioinformática podem utilizar a mineração de dados, extraindo de informações relevantes que auxiliam na elucidação de mecanismos moleculares de doenças (Fernández-Suárez and Birney, 2008). A enorme quantidade de dados brutos provenientes de estudos genéticos levou ao desenvolvimento de repositórios digitais, sítios onde os dados pudessem ser armazenados e disponibilizados para análises posteriores (Brazma, 2009).

Na genética humana, o genoma é formado pelo DNA, enquanto a expressão desses genes leva a produção de RNA, que pode codificar proteínas. O armazenamento de dados de expressão gênica (RNA), provenientes de estudos de microarray e sequenciamento de RNA (RNA-Seq), se dá, principalmente, em três bancos de dados públicos: Gene Expression Omnibus (GEO), ArrayExpress (AE) e Genomic Expression Archive (GEA). Apesar da centralização dessas informações nos repositórios citados, existem ainda dificuldades na realização de estudos nos metadados da pesquisa, devido à falta de padronização do vocabulário empregado pelos autores na estruturação dos dados armazenados (Wang et al., 2019).

Uma análise adequada das bases de dados de microarranjos pode melhorar a nossa compreensão da biologia e da medicina (Brazma, 2009). Apesar da enorme quantidade de dados transcriptômicos disponíveis nessas bases, há infelizmente uma subutilização dos mesmos (Huerta et al., 2014). O uso da tecnologia na análise dos dados a partir dessas bases torna o processo mais eficiente, possibilitando que o cientista tenha seu foco nos dados, e não na carga de trabalho necessária para analisa-los.

Grande parte dos algoritmos para análise de dados vindos das bases de dados de microarranjos são escritos na linguagem de programação R. No entanto, na análise desses dados, a utilização de linguagens de programação, na ausência de uma interface gráfica de utilização, se apresenta como uma barreira de dificuldade à comunidade de pesquisa em ciências da vida. Mesmo existindo alternativas como o Bioconductor, repositório que disponibiliza diversas ferramentas para esse fim (Nie et al., 2009), a falta de habilidades em programação pode afastar alguns pesquisadores dessas análises.

Visando amenizar as dificuldades associadas às presentes ferramentas de análise de dados transcriptômicos, o presente trabalho propõe o desenvolvimento de um sistema WEB capaz de ler dados brutos de pesquisas armazenadas no banco de dados biológico GEO, pré-processá-los, aplicar filtros para a seleção dos genes de interesse, e exibir ao usuário essas informações, juntamente de cálculos estatísticos e a expressão diferencial de cada gene selecionado, em uma interface acessível e usável.

O trabalho está organizado da seguinte forma: A Seção 2 apresenta a fundamentação teórica associada à temática desta pesquisa. A Seção 3 detalha a arquitetura do Mining_RNA, As Características do sistema e a validação. Por fim, a Seção 4 apresenta as conclusões e os possíveis trabalhos futuros.

2 Fundamentação teórica

A crescente quantidade de dados brutos de estudos genéticos sendo gerados torna imperiosa a tarefa de processá-los manualmente. A tarefa de combinação dos dados brutos da análise de microarranjo, de forma não automatizada, é inviável. O uso da bioinformática no processamento desta grande quantidade de dados, podendo também utilizar-se de técnicas de mineração de dados, viabiliza essas análises, possibilitando extrair novas informações a partir de estudos pré-existentes (Gangwar et al., 2012).

A mineração de dados utiliza algoritmos na busca por padrões válidos e compreensíveis dentro dos dados disponíveis, constituindo-se das seguintes etapas: associação, agrupamento e descoberta de regras de classificação (Espindola et al., 2010). Sua utilização é fundamental em bioinformática, por possibilitar que os pesquisadores passem a explorar as informações contidas nos dados (Fernández-Suárez and Birney, 2008), não mais exercendo o papel apenas de observadores das plataformas (por exemplo, ENSENBL e UCSC Genome Browser).

O uso de mineração ajuda na interpretação de dados que não são facilmente compreensíveis por meio de uma visualização gráfica, podendo também possibilitar a inferência de informações novas e úteis à descoberta de novos resultados de pesquisas novas ou pré-existentes (Garg et al., 2017). Os novos resultados podem ser alcançados por meio desta nova abordagem dos pesquisadores aos dados previamente analisados em outro estudo. As presentes dificuldades de análise dos conjuntos de dados depositados em bancos de dados biológicos públicos resulta em um aparente esquecimento dos mesmos, e novas análises podem proporcionar avanços em suas respectivas áreas de pesquisa (Lan et al., 2018).

Uma das possíveis maneiras de realizar mineração de dados biológicos é por meio dos recursos disponíveis no *Bioconductor* (Nie et al., 2009), um projeto de software livre que objetiva promover a análise estatística e o entendimento de dados de estudos biológicos. O projeto se baseia em pacotes escritos, majoritariamente, em linguagem de programação R, podendo conter contribuições de outras linguagens. O *Bioconductor* é um dos repositórios mais relevantes de ferramentas para o estudo de dados biológicos, disponibilizando pacotes destinados a diferentes tipos de análises, entre elas, a mineração de dados.

Embora exista uma diversidade de ferramentas úteis à análise de estudos biológicos, pesquisadores relatam certa dificuldade na utilização de algumas delas, frequentemente pela ausência de uma interface gráfica que possibilite o usuário a inserção de parâmetros para a análise dos estudos. Uma alternativa viável seria a utilização de um software ou sistema WEB que facilite esta análise, provendo uma interface gráfica de usabilidade. Os sistemas listados a seguir apresentam semelhanças ao software proposto por esta pesquisa, com o objetivo de captura e análise de dados de estudos disponíveis na base de dados biológicos GEO.

Chen et al. (2019) desenvolveram uma ferramenta a

qual denominaram de Restructured GEO (ReGEO), observando que os bancos de dados biológicos não armazenavam de forma estruturada os dados dos estudos, dificultando a reutilização e pesquisa desses dados. O ReGEO apresenta uma interface mais amigável ao pesquisador, se comparado com o GEO. Os desenvolvedores da ferramenta utilizaram técnicas de mineração de texto, visando analisar os metadados disponibilizados no GEO. A ferramenta 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 não anteriormente indexada. Disponibiliza-se também a busca por palavras-chave, como na ferramenta original. A abordagem utilizada pelos autores atingiu, de forma totalmente automática, uma taxa de precisão na mineração de texto de 93,5%.

Toro-Domínguez et al. (2018) apresentaram em sua pesquisa o desenvolvimento da plataforma ImaGEO. Uma ferramenta desenvolvida com base em Shiny, uma biblioteca designada a implementar uma interface WEB em conjunto com a linguagem de programação R. A solução apresentada conta com um interessante sistema dividido em 5 módulos. Os dados são carregados e processados utilizando um conjunto de parâmetros, que podem ser personalizados pelo usuário. Para o carregamento dos dados, utiliza-se o pacote GEOquery desenvolvido por Davis and Meltzer (2007). Outros módulos implementados pela ferramenta apresentam as seguintes finalidades: controle de qualidade, metanálise do *dataset*, análise funcional e por fim o módulo responsável pelo relatório. Este último abrange o panorama dos parâmetros utilizados na análise, um resumo relacionado ao processamento dos dados que identifica, por exemplo, *outliers* e valores ausentes. Em conclusão, também apresenta os resultados do processamento de dados, incluindo gráficos de calor para uma representação mais completa das informações obtidas sobre os genes. Apesar do ImaGEO ser uma ferramenta satisfatória em sua função, com uma interface gráfica intuitiva, existem ainda, no entanto, lacunas a serem preenchidas no que diz respeito à visualização final dos dados.

Uma aplicação de teor similar é o ScanGEO, desenvolvido por Koeppen et al. (2017). Desenvolvida em R, também utiliza o pacote shiny em sua implementação. A ferramenta permite, conforme os parâmetros preenchidos pelo usuário, consultas aos dados do GEO identificando estudos relevantes. De acordo com os autores, a aplicação dá suporte à análise da matriz de expressão gênica dos 20 principais organismos encontrados nas bases de dados do GEO, assim como buscas por palavras-chave. Na realização da pesquisa, o usuário poderá exportar os dados brutos obtidos para arquivos CSV para análise local ou em outros sistemas relacionados. Ainda segundo o autor, a ferramenta tem potencial considerável de acelerar a análise de dados públicos para a geração de hipóteses e validação de dados experimentais (Koeppen et al., 2017).

Djordjevic et al. (2019) propõem em seu artigo o GEOracle. Ferramenta que se utiliza de técnicas de mineração de texto e aprendizado de máquina para identificar, de maneira automática, experimentos de perturbação (ex. nocaute gênico), separar os diferentes grupos (experimental e controle), e avaliar a expressão diferencial. O sistema apresentado é uma ferramenta de código livre, que apesar do autor não informar um endereço onde a ferramenta pode ser acessada diretamente, disponibiliza o código-fonte na plataforma de repositórios GitHub. Apesar de oferecer uma interface relativamente intuitiva, conta com menos funcionalidades se comparada com o ImaGEO.

O GEO2R se trata de uma aplicação mantida pelo NCBI que possibilita que os usuários comparem dois ou mais grupos de amostras a partir de estudos armazenados no GEO. O sistema permite a identificação da expressão diferencial dos genes de um estudo, e retorna os dados em formato de tabela com os genes por ordem de significância, além de gráficos que auxiliam na visualização dos genes diferencialmente expressos (NCBI, 2021a).

A interface online do GEO2R possibilita a análise dos 250 genes mais relevantes. Para visualizar o restante dos genes, o sistema permite o download da tabela para carregamento em outros programas, como o Microsoft Excel ou semelhante. A aplicação retorna ainda um código em R que pode ser executado no computador do pesquisador. No entanto, para conseguir alterar todos os parâmetros, necessita-se o conhecimento prévio na linguagem de programação pelo pesquisador.

Os sistemas supracitados objetivam prioritariamente facilitar a análise de dados para os pesquisadores. Majoritariamente, tratam-se de ferramentas que leem parte dos dados que a plataforma GEO não expõe em sua interface, tendo em vista que algumas destas informações não estão devidamente alocadas, e sim descritas no contexto do estudo. Outras ferramentas conseguem ligar-se 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 a visualização dos dados obtidos a partir de uma série de filtros direcionados a 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. Estas foram as funcionalidades implementadas no sistema desenvolvido pelo presente trabalho.

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

O sistema Mining_RNA visa facilitar a mineração de dados dos estudos transcriptômicos depositados no banco de dados GEO. Ele 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 estatística (teste-t e valor-p) para cada gene do estudo analisado. A seguir serão apresentadas a arquitetura, as características e a validação do sistema.

3.1 Arquitetura do sistema Mining_RNA

A Fig. 1 apresenta a arquitetura do sistema, que consiste em 3 API's, um banco de dados e uma interface WEB que

permite que o usuário realize consultas aos dados. A divisão entre API's e interfaces se deve à possibilidade de alocar os algoritmos em diferentes servidores. Alguns dos algoritmos que compõem as API's do sistema podem ser computacionalmente complexos, podendo resultar em uma carga elevada de processamento. A arquitetura descentralizada apresenta-se como uma alternativa mais eficiente para melhor escalonamento.

A API 01 executa as tarefas de mineração dos dados obtidos, assim como as tarefas de aprendizagem de máquina. Os algoritmos implementados nessa API possibilitam a validação dos dados antes do retorno ao usuário. Objetivando a validação interna, o algoritmo *Random Forest* foi utilizado. Seu uso se justifica por sua adequação quando o número de atributos é maior que o número de amostras, em um conjunto de dados, sendo esta uma forte característica dos microarranjos de DNA. Esta mesma API poderá também dar suporte à implementação de outros algoritmos de mineração, em futuros trabalhos que estenderão essa pesquisa.

A API 02 realiza o pré-processamento dos dados carregados no sistema WEB,podendo ainda fazer uso de dados diretamente do banco de dados local, assim como solicitar que a API 03 faça busca de novos dados no GEO. Esta API foi desenvolvida respeitando rigorosamente princípios de reúso, para que novos filtros sejam inclusos em futuras iterações do projeto.

A API 03 é responsável por capturar dos dados do *dataset* solicitado pelo usuário e armazenar as informações relevantes no banco de dados local. Apesar da simplicidade da tarefa, por conta da complexidade dos dados apresentados por algumas pesquisas, ocasionalmente é necessário grande poder de processamento. A API interpreta na íntegra esses dados, e os armazena, para acelerar etapas posteriores de processamento. As três API's foram desenvolvidas utilizando a linguagem de programação Python.

O sistema WEB consiste na interface na qual o usuário realiza suas requisições. Esta parte do sistema foi desenvolvida na linguagem de programação PHP para facilitar sua disponibilização online. Esta interface foi construída objetivando a facilidade de operação, evitando dessa forma que filtros importantes sejam deixados de lado por sua complexidade. A Fig. 2 demonstra os filtros aplicados na versão inicial do sistema. Trata-se de uma interface acessível e usável com mecanismos de ajuda para o usuário. Pretende-se que seja um sistema adaptável, embora, momentaneamente, não esteja no escopo deste trabalho uma interface específica para dispositivos com telas de baixa resolução. Esta decisão se justifica pela densidade dos dados a serem exibidos, necessitando, por enquanto, de uma área de tela maior para uma visualização inteligível.

É necessário que os usuários do sistema estejam devidamente cadastrados para que possa haver uma personalização adequada, assim como a realização de análises relevantes para a tomada de decisão relacionada à alocação de recursos, e implementações futuras. Essa autenticação pretende também prevenir erros relacionados à usabilidade, visando preservar a disponibilidade e integridade do sistema.

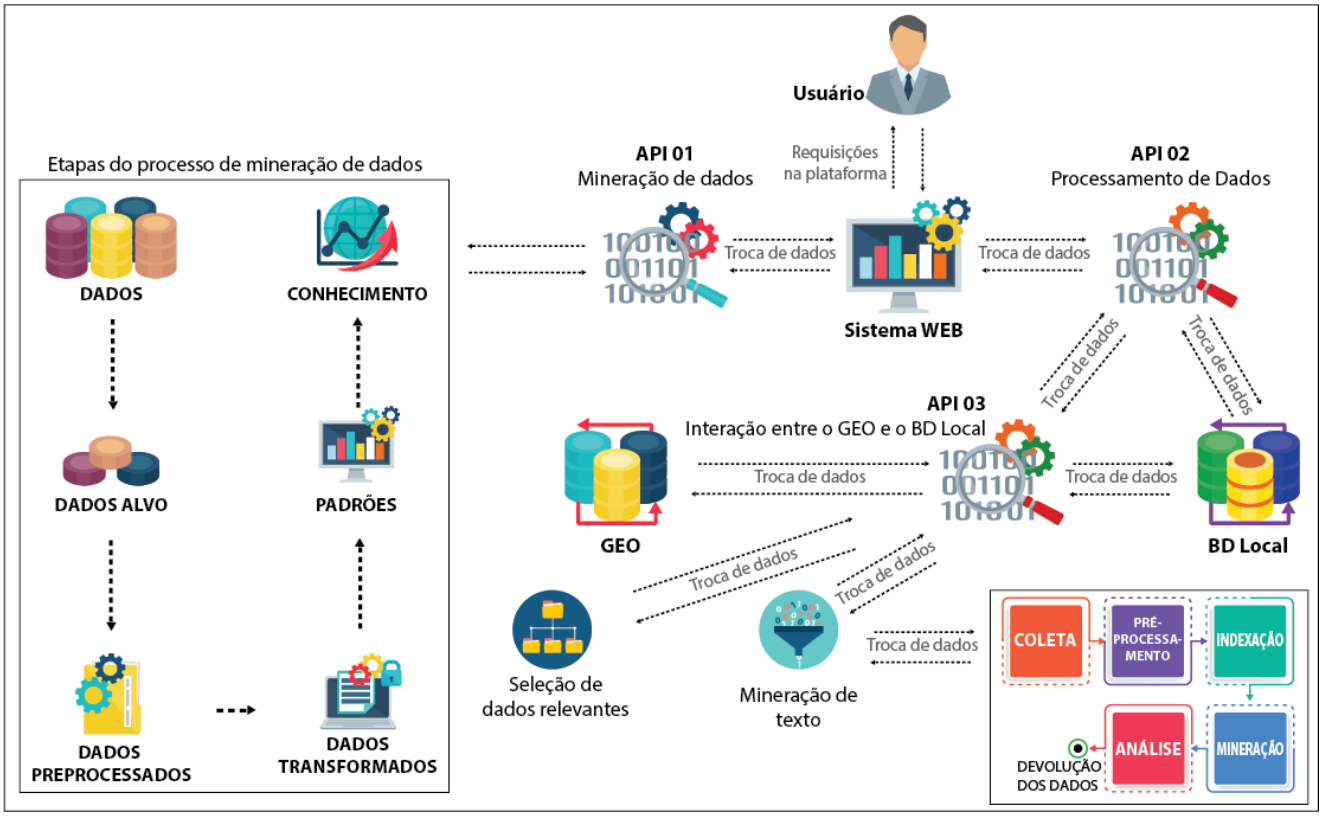


Figura 1: Visão geral da arquitetura do sistema Mining_RNA. Fonte: Adaptado de Moreira et al. (2021)

Por fim, a base de dados relacional, que armazena os dados relevantes para as consultas, permite o armazenamento de dados processados, possibilitando ao usuário recuperá-los em um momento futuro, armazenando também *presets* de filtros definidos pelo usuário para suas respectivas aplicações em situações futuras.

3.2 Características do sistema desenvolvido

Partindo da arquitetura planejada, também descrita em Moreira et al. (2021), foi desenvolvido o Mining_RNA: um sistema web capaz de reanalisar os dados brutos de estudos biológicos. Para a análise dessas pesquisas pelo sistema, o usuário deverá percorrer quatro passos: Seleção do estudo, escolha dos grupos para análise, configuração de filtros e, exibição dos resultados.

Inicialmente, o código do estudo proveniente do banco de dados biológico GEO é inserido em uma caixa de busca no sistema. Em seguida, o sistema analisa os metadados do estudo, e retorna uma tela de seleção do grupo de controle e o conteúdo dos casos que serão comparados. Depois da devida seleção dos grupos, o próximo passo possibilita a personalização de parâmetros de filtragem dos dados do estudo, para a relevância dos dados trazidos no resultado, como demonstra a Fig. 2. A opção "Ponto de corte FC"limita 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

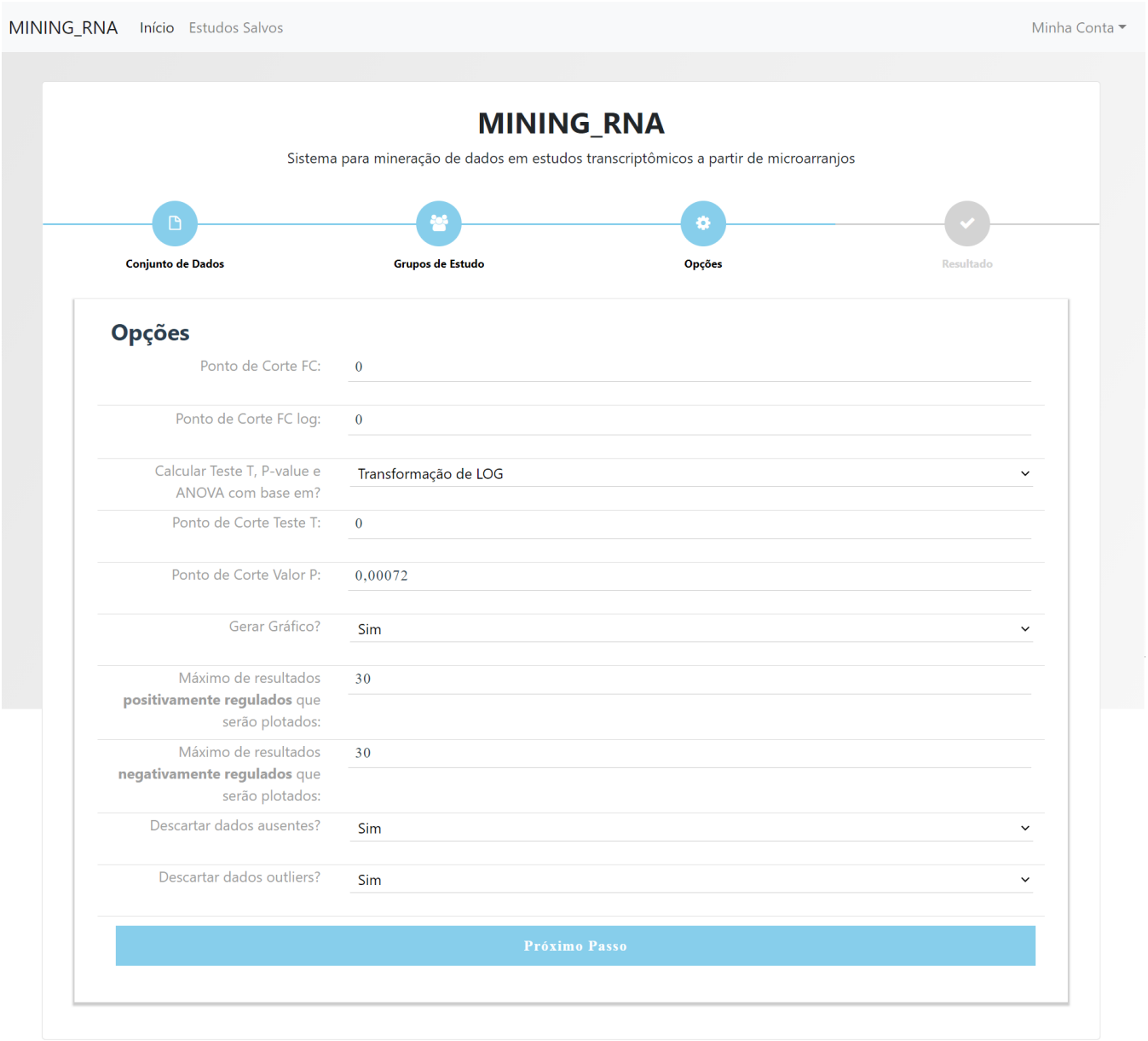


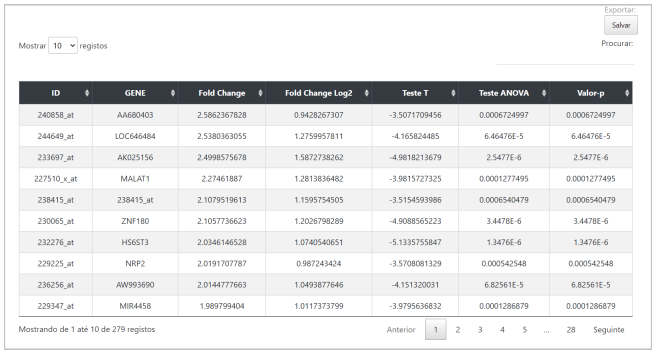
Figura 2: Tela de filtros do sistema Mining_RNA

comparação das amostras do grupo de controle em relação às amostras do grupo de casos. O cálculo de fold change demonstra, quantitativamente, a expressão de determinado gene em diferentes situações. É possível, também, definir um ponto de corte para o fold change, que foi calculado utilizando a base logarítmica 2, pela opção "Ponto de corte FC Log".

Para que também seja possível analisar os dados estatisticamente, 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, uma vez que cada grupo é composto de várias amostras. Devido a isso, foi preciso o cálculo da média dos valores de cada gene dessas amostras. Devido à comparação das médias, é importante a realização deste teste para a confiabilidade dos dados. O cálculo do valor P se utiliza em diversos cálculos estatísticos, inclusive nos Testes T, que facilitam verificações realizadas pelo pesquisador no sistema. Ambos estes valores podem ter sua exibição limitada pelas opções "Ponto de corte Teste T"e "Ponto de Corte Valor P".

O sistema também pode retornar ao usuário um gráfico relacionado ao cálculo de *fold change*. Essa opção pode também ser definida nos controles ilustrados na Fig. 2. Nessa opção, o pesquisador define a quantidade de resultados sendo exibida para genes positivamente regulados e negativamente regulados. Nessa mesma 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.

Após a seleção dos filtros a serem aplicados, o sistema disponibiliza dois tipos de visualização dos dados calculados. A primeira exibe uma tabela, conforme demonstrado na Fig. 3. A tabela contém 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, há uma ordenação dos dados pelo valor do *fold change*. No entanto, a interatividade da tabela possibilita a ordenação os dados por qualquer uma das colunas disponibilizadas, podendo ainda aumentar a quantidade de registros exibidos em cada página, e realizar uma pesquisa em qualquer dado disposto na tabela.


Exportar:
Salvar
Mostrar 10 registos
Procurar:

ID	GENE	Fold Change	Fold Change Log2	Teste T	Teste ANOVA	Valor-p
240858_at	AA680403	2.5862367828	0.9428267307	-3.5071709456	0.0006724997	0.0006724997
244649_at	LOC646484	2.5380363055	1.2759957811	-4.165824485	6.46476E-5	6.46476E-5
233697_at	AK025156	2.4998575678	1.5872738262	-4.9818213679	2.5477E-6	2.5477E-6
227510_x_at	MALAT1	2.27461887	1.2813836482	-3.9815727325	0.0001277495	0.0001277495
238415_at	238415_at	2.1079519613	1.1595754505	-3.5154593986	0.0006540479	0.0006540479
230065_at	ZNF180	2.1057736623	1.2026798289	-4.9088565223	3.4478E-6	3.4478E-6
232276_at	HS6ST3	2.0346146528	1.0740540651	-5.1335755847	1.3476E-6	1.3476E-6
229225_at	NRP2	2.0191707787	0.987243424	-3.5708081329	0.000542548	0.000542548
236256_at	AW993690	2.0144777663	1.0493877646	-4.151320031	6.82561E-5	6.82561E-5
229347_at	MIR4458	1.989799404	1.0117373799	-3.9795636832	0.0001286879	0.0001286879

Mostrando de 1 até 10 de 279 registos
Anterior 1 2 3 4 5 ... 28 Seguinte

Figura 3: Visão geral da tela de resultados - Parte 1 (Tabela)

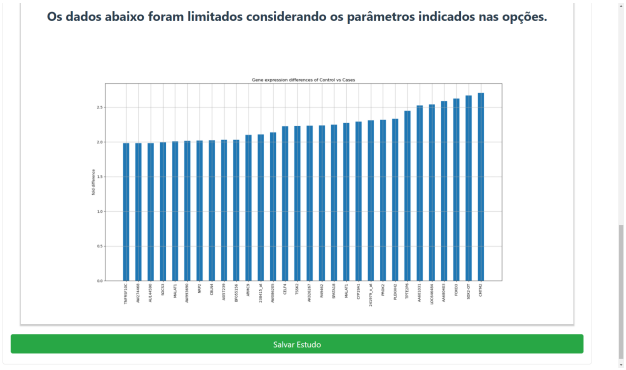


Figura 4: Visão geral da tela de resultados - Parte 2 (Gráfico)

É importante para o pesquisador que o sistema possibilite a reutilização dos dados obtidos em outras análises. Devido esta necessidade, é disponibilizado juntamente à tabela de resultados um botão que ativa a função de exportação dos dados ali dispostos para um arquivo compatível com um editor de planilhas convencional, como o Microsoft Excel ou o OpenOffice Calc. Ao selecionar essa opção, é realizado o download dos dados da tabela para o computador do usuário. Na posse desses dados, o pesquisador pode adaptar o formato das informações para torná-las compatíveis com outras aplicações de análise genética, dando assim ainda mais profundidade à sua pesquisa.

A segunda forma que o sistema disponibiliza os dados de *fold Change* para análise específica é por meio do gráfico, como ilustrado na Fig. 4. O gráfico é mostrado ao usuário na forma de uma imagem que, além de visível na tela, pode ser salva pelo pesquisador seguindo os devidos passos padrão do próprio navegador sendo utilizado. Este tipo de representação dos dados pode ajudar no entendimento das informações ali dispostas, ou ainda auxiliar que o pesquisador possa enviá-las a terceiros ou utilizá-las em apresentações dos resultados obtidos.

3.3 Validação

Para a execução dos testes de validação do sistema Mining_RNA,foi selecionado um estudo depositado na base de dados biológicos GEO. O estudo está identificado pelo código GSE9006, e analisa 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 este objetivo, foram realizadas comparações utilizando o próprio artigo publicado a partir dos dados do GSE9006, junto de análises utilizando outra ferramenta disponibilizada pelo próprio NCBI, o GEO2R.

Com objetivo de selecionar um subconjunto de genes significativos, foi definido que o ponto de corte do valor-p seria de 0,00072. Este ponto de corte foi escolhido por ser também um dos parâmetros utilizados na pesquisa original. A partir desse fator limitante, foram realizados testes utilizando os dados brutos obtidos através de microarranjo, junto desses mesmos valores após a aplicação da transfor-

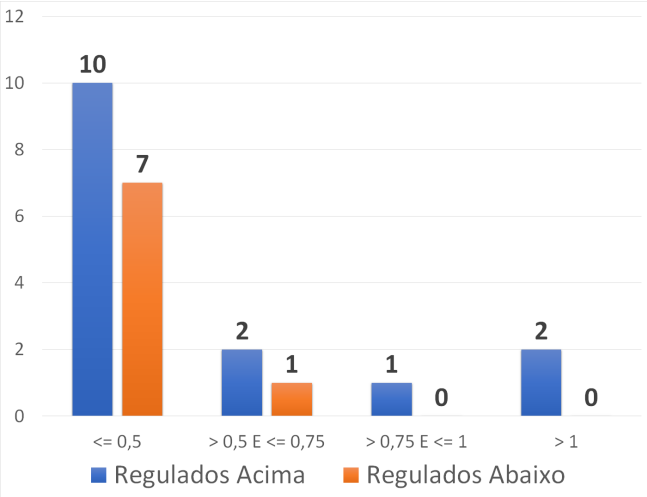


Figura 5: Diferença entre valores de *fold change* obtidos pelo Mining_RNA e os encontrados no estudo original

mação para log2. A análise com os dados transformados se faz necessária pelo fato de que as bibliotecas consolidadas em R para análises biológicas, como o *limma*, necessitam que os valores sejam tratados desta maneira para realizar seus cálculos de *fold change*.

As análises realizadas a partir dos dados do estudo GSE9006 mostraram que o nível de eficiência apresentado pelo Mining_RNA é satisfatório. A Tabela 1 sintetiza as eficiências obtidas nos três testes realizados para a validação do sistema.

Tabela 1: 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%

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

O terceiro teste, realizado pela comparação dos resultados calculados a partir do objeto produzido por esta pesquisa junto do 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 Fig. 5. Este resultado proporciona a uma validação satisfatória para os resultados obtidos nesta pesquisa, mostrando ainda que os as informações obtidas através do sistema são confiáveis, pela aproximação às obtidas através de outros tipos de estudos já publicados em meios científicos.

4 Considerações Finais

Este trabalho apresenta o Mining_RNA, um sistema web para mineração de dados em estudos transcriptômicos a partir de microarranjos, com interface desenvolvida objetivando a facilidade de utilização. O sistema permite que os pesquisadores realizem reanálises em estudos armazenados no banco de dados biológicos GEO.

Por meios de poucos passos, poucas interações e a seleção de filtros, o sistema proporciona aos pesquisadores a oportunidade de reanalisar resultados de aproximadamente 143 mil (NCBI, 2021b) estudos transcriptômicos, sem a necessidade de conhecimento prévio em programação, promovendo um ganho de tempo que pode ser dedicado à análise final dos resultados apresentados pelo sistema.

Para a confirmação da satisfatoriedade dos resultados apresentados no sistema desenvolvido, foi realizada a validação dos mesmos em comparação com valores gerados através da ferramenta GEO2R, do NCBI, e 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.

Perspectivas futuras no desenvolvimento de novas versões sistema Mining_RNA, poderão incluir as seguintes como 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.

Agradecimentos

Os autores agradecem ao Grupo de Engenharia e Software da Universidade do Estado do Rio Grande do Norte, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento das bolsas aos pesquisadores.

Referências

Brazma, A. (2009). Minimum information about a microarray experiment (miame) – successes, failures, challenges, *TheScientificWorldJournal* **9**: 420–3. https://doi.org/10.1100/tsw.2009.57.

Chen, G., Ramírez, J. C., Deng, N., Qiu, X., Wu, C., Zheng, W. J. and Wu, H. (2019). Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis, *Database* **2019**. https://doi.org/10.1093/database/bay145.

Davis, S. and Meltzer, P. S. (2007). Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor, *Bioinformatics* **23**(14): 1846–1847. https://doi.org/10.1093/bioinformatics/btm254.

Djordjevic, D., Tang, J. Y., Chen, Y. X., Kwan, S. L. S., Ling, R. W., Qian, G., Woo, C. Y., Ellis, S. J. and Ho, J. W. (2019). Discovery of perturbation gene targets via free text metadata mining in gene expression omnibus, *Computational Biology and Chemistry* **80**: 152 – 158. https://doi.org/10.1016/j.compbiolchem.2019.03.014.

Espindola, F. S., Calábria, L. K., de Rezende, A. A. A., Pereira, B. B., Santana, F. A., Amaral, I. M. R., Lobato, J., França, J. L., Mario, J. L., Figueiredo, L. B. et al. (2010). Recursos de bioinformática aplicados às ciências ômicas como genômica, transcriptômica, proteômica, interatômica e metabolômica, *Bioscience Journal* **26**(3). Disponível em https://pesquisa.bvsalud.org/portal/resource/pt/lil-561959.

Fernández-Suárez, X. M. and Birney, E. (2008). Advanced genomic data mining, *PLoS computational biology* **4**(9): e1000121–e1000121. https://doi.org/10.1371/journal.pcbi.1000121.

Gangwar, V., Ghose, U. and Singh, Y. (2012). 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* **3**(3). https://doi.org/10.26483/ijarcs.v3i3.1178.

Garg, S. B., Mahajan, A. K. and Kamal, T. (2017). An approach for diabetes detection using data mining classification techniques, *Journal of Engineering Sciences* **26**. Disponível em http://ijoes.vidyapublications.com/paper/Vol26/20-Vol26.pdf.

Gasparovica-Asīte, M. and Aleksejeva, L. (2019). Classification methodology for bioinformatics data analysis, *Automatic Control and Computer Sciences* **53**(1): 28–38. https://doi.org/10.3103/S0146411619010073.

Huerta, M., Munyi, M., Expósito, D., Querol, E. and Cedano, J. (2014). Mgdb: Crossing the marker genes of a user microarray with a database of public-microarrays marker genes, *Bioinformatics (Oxford, England)* **30**. https://doi.org/10.1093/bioinformatics/btu109.

Kaizer, E. C., Glaser, C. L., Chaussabel, D., Banchereau, J., Pascual, V. and White, P. C. (2007). Gene Expression in Peripheral Blood Mononuclear Cells from Children with Diabetes, *The Journal of Clinical Endocrinology & Metabolism* **92**(9): 3705–3711. https://doi.org/10.1210/jc.2007-0979.

Koeppen, K., Stanton, B. A. and Hampton, T. H. (2017). ScanGEO: parallel mining of high-throughput gene expression data, *Bioinformatics* **33**(21): 3500–3501. https://doi.org/10.1093/bioinformatics/btx452.

Lan, K., Wang, D.-t., Fong, S., Liu, L.-s., Wong, K. K. and Dey, N. (2018). A survey of data mining and deep learning in bioinformatics, *Journal of medical systems* **42**(8): 139. https://doi.org/10.1007/s10916-018-1003-9.

Moreira, C. R., Pacheco, C., Diógenes, M. V. P., Batista, P. V. M., Neto, P. F. R., Silva, A. G. d., Felipe, S. M. d. S., Ceccatto, V. M., Freitas, R. M. d., Gurgel, T. K. S., Santos, E. C. d., Maia, C. M. and Leite, T. A. A. e. C. R. M. S. (2021). Mining_rna: Web-based system using e-science for transcriptomic data mining, *in* H. R. Arabnia, L. Deligiannidis, M. R. Grimaila, D. D. Hodson, K. Joe, M. Sekijima and F. G. Tinetti (eds), *Advances in Parallel & Distributed Processing, and Applications*, Springer International Publishing, Cham, pp. 1195–1203. https://doi.org/10.1007/978-3-030-69984-0_85.

NCBI (2021a). About geo2r, *Technical report*, National Center for Biotechnology Information. Disponível em https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html.

NCBI (2021b). Geo - ncbi, *Technical report*, National Center for Biotechnology Information. Disponível em https://www.ncbi.nlm.nih.gov/geo/.

Nie, H., Neerincx, P., van der Poel, J., Ferrari, F., Bicciato, S., Leunissen, J. and Groenen, M. (2009). Microarray data mining using bioconductor packages, *BMC Proceedings* **3**(Suppl.4): S9. https://doi.org/10.1186/1753-6561-3-S4-S9.

Toro-Domínguez, D., Martorell-Marugán, J., López-Domínguez, R., García-Moreno, A., González-Rumayor, V., Alarcón-Riquelme, M. E. and Carmona-Sáez, P. (2018). ImaGEO: integrative gene expression meta-analysis from GEO database, *Bioinformatics* **35**(5): 880–882. https://doi.org/10.1093/bioinformatics/bty721.

Wang, Z., Lachmann, A. and Ma'ayan, A. (2019). Mining data and metadata from the gene expression omnibus, *Biophysical Reviews* **11**(1): 103–110. https://doi.org/10.1007/s12551-018-0490-8.