UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

FABIANO BORGES PEREIRA

DESENVOLVIMENTO DE UM SOFTWARE BIOINFORMÁTICO ESCRITO EM LINGUAGEM DE PROGRAMAÇÃO PYTHON PARA LEITURA E EXTRAÇÃO DE INFORMAÇÕES DE ARQUIVOS PROVENIENTES DO GEOCANCERPROGNOSTICDATASETSRETRIEVER

PATOS DE MINAS-MG
2023

FABIANO BORGES PEREIRA

DESENVOLVIMENTO DE UM SOFTWARE BIOINFORMÁTICO ESCRITO EM LINGUAGEM DE PROGRAMAÇÃO PYTHON PARA LEITURA E EXTRAÇÃO DE INFORMAÇÕES DE ARQUIVOS PROVENIENTES DO GEOCANCERPROGNOSTICDATASETSRETRIEVER

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Biotecnologia como requisito parcial para a obtenção do título de Mestre em Biotecnologia.

Orientador: Prof. Dr. Laurence Rodrigues do Amaral

PATOS DE MINAS

2023

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

P436 Pereira, Fabiano Borges, 1979-
2023 DESENVOLVIMENTO DE UM SOFTWARE BIOINFORMÁTICO ESCRITO EM LINGUAGEM DE PROGRAMAÇÃO PYTHON PARA LEITURA E EXTRAÇÃO DE INFORMAÇÕES DE ARQUIVOS PROVENIENTES DO GEOCANCERPROGNOSTICDATASETSRETRIEVER [recurso eletrônico] / Fabiano Borges Pereira. - 2023.

Orientador: Laurence Rodrigues do Amaral.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Biotecnologia.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2024.4 Inclui bibliografia.

1. Biotecnologia. I. Amaral, Laurence Rodrigues do, 1978-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Biotecnologia. III. Título.

CDU: 60

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Biotecnologia
Av. Getúlio Vargas, 230, 3º andar, Sala 308 - Bairro Centro, Patos de Minas-MG, CEP 38700-128
Telefone: (34) 3823-3714 - Ramal 39 - www.ppgbiotec.ibtec.ufu.br - ppgbiotec@ibtec.ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Biotecnologia				
Defesa de:	Dissertação de Mestrado Acadêmico PPGBIOTEC				
Data:	20/12/2023	Hora de início:	09:00	Hora de encerramento:	12:38
Matrícula	42112BTC003				
Nome do Discente:	Fabiano Borges Pereira				
Título do Trabalho:	Desenvolvimento de um Software Bioinformático Escrito em Linguagem de Programação Python para Leitura e Extração de Informações de Arquivos Provenientes do geoCancerPrognosticDatasetsRetriever				
Área de concentração:	Biociências				
Linha de pesquisa:	Bioinformática e Biologia Molecular aplicada à genômica, transcriptômica e proteômica				
Projeto de Pesquisa de vinculação:					

Reuniu-se remotamente, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Biotecnologia, assim composta: Professores Doutores: Letícia Conceição Braga - Instituto Mario Penna; Pedro Luiz Lima Bertarini FEELT/UFU; Laurence Rodrigues do Amaral - FACOM /UFU, orientador do candidato.

Iniciando os trabalhos o presidente da mesa, Dr. Laurence Rodrigues do Amaral, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu o Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Laurence Rodrigues do Amaral**, **Professor(a) do Magistério Superior**, em 20/12/2023, às 16:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Letícia da Conceição Braga**, **Usuário Externo**, em 21/12/2023, às 06:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Pedro Luiz Lima Bertarini**, **Professor(a) do Magistério Superior**, em 21/12/2023, às 08:10, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5050513** e o código CRC **84065F12**.

Referência: Processo nº 23117.088853/2023-58 SEI nº 5050513

FABIANO BORGES PEREIRA

DESENVOLVIMENTO DE UM ALGORITMO BIOINFORMÁTICO EM PYTHON PARA LEITURA E EXTRAÇÃO DE DADOS DE ARQUIVOS PROVENIENTES DO SOFTWARE GEOCANCERPROGNOSTICDATASETSRETRIEVER

Com honra e dedicação, submeto à apreciação do Programa de Pós-graduação em Biotecnologia, a presente Dissertação de Mestrado como requisito parcial para a obtenção do título de Mestre em Biotecnologia.

APROVADO EM: ____ / ____ / ____

BANCA EXAMINADORA

_____

Prof. Dr. Laurence Rodrigues do Amaral

_____

Prof. Dr. Pedro Luiz Lima Bertarini

_____

Profa. Dra. Letícia Conceição Braga

PATOS DE MINAS – MG
2023

GRATIDÕES

Sou extremamente grato ao meu Deus por me fazer criar casca, por me envolver em sua couraça e por me circundar com o seu fogo. E ao meu Guardião, que faz com que eu sinta que um amigo está ao meu lado, que vê meus sofrimentos e compartilha minhas alegrias.

Manifesto minha profunda gratidão à minha mãe, Dalva Borges, cujo apoio e dedicação garantiram não apenas as condições necessárias, mas também um suporte inestimável em todos os aspectos, permitindo-me concluir com sucesso este trabalho tão importante.

A mim, por suportar e ser muito melhor que a toxina de pessoas que por alguma motivação, tentaram me atingir negativamente.

Sou grato à minha amiga Sarah Oliveira, que sempre acreditou em mim, sempre esteve do meu lado e sempre me ajudou em vários sentidos.

Sou grato à minha amiga Juliana Dias, que sempre torceu por mim e sempre esteve pronta para me ajudar no que fosse preciso e possível.

Sou grato à FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) que tem a nobre missão de estimular e fomentar o desenvolvimento científico e tecnológico no estado e que acreditou em minha pesquisa e me incentivou por um ano.

Dedico este trabalho a todos os corajosos pacientes que enfrentam a luta contra o câncer dia após dia. Dedico também aos pesquisadores comprometidos que se dispõem a avançar no campo da Bioinformática e da Biotecnologia, utilizando o código UNITE-1 como um ponto de partida para criar novos algoritmos ainda mais eficientes e capazes de melhorar a compreensão da expressão gênica.

Que este trabalho seja um tributo à esperança e ao progresso na luta contra essa estranha doença.

“Da inspiração ao algoritmo, do algoritmo à sapiência. Um novo código, uma nova chave.”

Fabiano Borges Pereira

RESUMO

A Bioinformática é fundamental para compreender os mais diferentes processos biológicos, principalmente a nível molecular. Assim, nesse trabalho é apresentado uma ferramenta de busca chamada geoCancerPrognosticDatasetsRetriever como uma solução para diminuir as dificuldades dos pesquisadores em recuperar com eficiência e rapidez os conjuntos de dados de expressão gênica ligados ao prognóstico de câncer, que estão disponíveis no banco de dados do GEO (*Gene Expression Onminus*). Porém, mesmo a ferramenta fazendo essa otimização, a quantidade de informações em cada arquivo ainda é muito grande, o que dificulta o processamento manual. Para resolver esse problema foi criado o UNITE-1 (*Understanding Novel Information Through Expression*) para agilizar a extração dos dados que estão relacionados principalmente ao prognóstico do câncer de algum arquivo disponível no GEO, gerando um novo arquivo personalizado, resolvendo com praticidade, rapidez e eficiência mais um importante problema enfrentado por pesquisadores que precisam analisar dados de expressão gênica para compor os seus trabalhos. Com a combinação dessas duas ferramentas, o geoCancerPrognosticDatasetsRetriever (*Perl*) e o UNITE-1 (*Python*), os pesquisadores podem analisar os dados relacionados ao prognóstico de algum tipo de câncer, com eficiência e rapidez, mostrando um potencial não apenas de impulsionar suas pesquisas científicas, mas também facilitando a criação de novas propostas para combater a doença.

Palavras-chave: Bioinformática, geoCancerPrognosticDatasetsRetriever, UNITE-1, Câncer, GEO.

ABSTRACT

Bioinformatics is fundamental to understanding many different biological processes, especially at the molecular level. Therefore, this paper presents a search tool called geoCancerPrognosticDatasetsRetriever as a solution to reduce the difficulties researchers have in efficiently and quickly retrieving gene expression datasets linked to cancer prognosis, which are available in the GEO (Gene Expression Onminus) database. However, even though the tool does this optimization, the amount of information in each file is still very large, which makes manual processing difficult. To solve this problem, UNITE-1 (Understanding Novel Information Through Expression) was created to speed up the extraction of data that is mainly related to cancer prognosis from any file available in GEO, generating a new personalized file, solving another important problem faced by researchers who need to analyze gene expression data to compose their work with practicality, speed and efficiency. With the combination of these two tools, geoCancerPrognosticDatasetsRetriever (Perl) and UNITE-1 (Python), researchers can delve into data related to the prognosis of some type of cancer, efficiently and quickly, showing the potential not only to boost their scientific research, but also to facilitate the creation of new proposals to combat the disease.

Keywords: Bioinformatics, geoCancerPrognosticDatasetsRetriever, UNITE-1, Cancer, GEO.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ADJ – *Adjuvant* / Adjuvante.

BIG – *Breast International Group* / Grupo Internacional da Mama.

DNA – *Deoxyribonucleic Acid* / Ácido Desoxirribonucleico.

ER – *Estrogen receptor* / Receptor de Estrogênio.

GEO – *Gene Expression Omnibus*.

GGI – *Gene Expression Grade Index* / Índice de Grau de Expressão Gênica.

GPL – *GEO Platform registry* / Registro da Plataforma GEO.

GSE – *GEO Series registry* / Registro da Série GEO.

GSM – *GEO sample registry* / Registro de Amostra GEO.

HER2 – *Human Epidermal Growth Factor Receptor* 2 / Receptor 2 do Fator de Crescimento Epidérmico Humano.

NEOADJ – *Neoadjuvant* / Neoadjuvante.

PGR – *Progesterone Receptor* / Receptor de Progesterona.

RFS – *Relapse-Free Survival* / Sobrevivência Livre de Recidiva.

RNA - *Ribonucleic Acid* / Ácido Ribonucleico.

T RFS – *Time to Relapse-Free Survival* / Tempo Para Sobrevida Livre de Recidiva.

UNITE - *Understanding Novel Information Through Expression* / Compreendendo Novas Informações por Meio da Expressão.

YT3 – *Anonymous Sample Group Identifier* / Identificador Anônimo de Grupo de Amostra.

SUMÁRIO

1 INTRODUÇÃO

A Bioinformática é uma área capaz de lidar com grandes volumes de dados biológicos a partir da combinação de diferentes áreas interdisciplinares como a Biologia, a Estatística e a Computação. A Bioinformática tem se tornado cada vez mais importante e indispensável para as pesquisas na Biologia e na Medicina por causa do acelerado crescimento tecnológico e pela grande disponibilidade de dados genômicos. O prognóstico e o diagnóstico do câncer, uma doença complexa e variada que afeta milhões de pessoas no mundo inteiro, podem ser melhor compreendidos pelo uso de dados genômicos que aumentam constantemente, à medida que as pesquisas são realizadas e publicadas. Assim, a criação de ferramentas e algoritmos computacionais para analisar os dados genômicos relacionados ao câncer é muito importante para a compreender melhor essas informações e a sua relação com os desfechos clínicos abordados, e então desenvolver formas de combate a essa doença.

Nesse trabalho, é apresentado duas ferramentas para auxiliar na pesquisa do prognóstico do câncer, o geoCancerPrognosticDatasetsRetriever e o UNITE-1. O geoCancerPrognosticDatasetsRetriever é uma ferramenta de busca que identifica no banco de dados do GEO, os conjuntos de dados de expressão gênica que tem informações sobre o prognóstico do câncer. O UNITE-1 é um programa de computador desenvolvido em linguagem Python, que tem como objetivo resolver o problema de extrair dos complexos e grandes arquivos do GEO obtidos pelo geoCancerPrognosticDatasetsRetriever, somente as informações relacionadas ao prognóstico do câncer.

O principal objetivo desse trabalho é desenvolver uma ferramenta para otimizar a busca por termos específicos e que estão relacionados ao prognóstico do câncer de interesse para poder facilitar a pesquisa e a análise das informações desejadas sobre essa doença. Para isso, utilizou-se primeiramente o geoCancerPrognosticDatasetsRetriever, que é uma ferramenta que busca no banco de dados do GEO (*Gene Expression Omnibus*) apenas arquivos de conjuntos de dados que tem informações sobre o prognóstico do câncer pesquisado. Mesmo com essa otimização, a quantidade de dados em cada arquivo é muito grande. Nesse contexto, foi desenvolvido, com a colaboração do modelo de linguagem de inteligência artificial GPT-4 criado pela OpenAI, o UNITE-1 (*Understanding Novel Information Through Expression*), um programa de computador escrito em linguagem de programação Python que processa os arquivos '.soft' que são gerados pela ferramenta geoCancerPrognosticDatasetsRetriever gerando um novo arquivo personalizado contendo apenas as informações desejadas que foram extraídas a partir das palavras-chave que foram previamente definidas. Com o intuito de avaliar

a eficiência do UNITE-1, foi conduzido testes utilizando conjuntos de dados relacionados aos cânceres colorretal e de mama, onde foi comprovado que o algoritmo é capaz de processar os arquivos do GEO com extensão ‘.soft’.

O UNITE-1 é executado no interpretador Jupyter Notebook sem a necessidade de estar conectado à internet. Esse estudo é uma contribuição importante para o vasto campo da Bioinformática, oferecendo uma ferramenta computacional para auxiliar os pesquisadores e os especialistas da área da saúde, que se dedicam a estudar as informações de expressão gênica do câncer, armazenadas no banco de dados do Gene Expression Onminus (GEO).

O presente trabalho está estruturado em seis seções principais. No capítulo 1 é apresentado uma introdução sobre o tema. No capítulo 2 aborda a fundamentação teórica, que inclui uma visão geral da bioinformática e os recursos do banco de dados do GEO (*Gene Expression Omnibus*). O capítulo 3 descreve em detalhes o desenvolvimento do programa de computador em Python UNITE-1 para análise de prognóstico de câncer. O capítulo 4 analisa a utilização e os resultados do UNITE-1. O capítulo 5 apresenta as conclusões e o capítulo 6 propõe possibilidades para trabalhos futuros.

O objetivo desse trabalho é desenvolver um programa de computador que analisa os resultados gerados pelo geoCancerPrognosticDatasetsRetriever, extrai as informações relacionadas ao prognóstico, e as organiza de acordo com cada amostra, maximizando a eficiência na análise dos resultados, otimizando o trabalho dos pesquisadores na busca de biomarcadores em oncologia.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta uma visão geral da bioinformática e o seu uso na análise de dados de prognóstico do câncer, abordando conceitos importantes como analisar dados biológicos a partir do uso de ferramentas computacionais. O capítulo discute como a bioinformática combina conhecimentos de diferentes áreas como a computação, a estatística e a biologia. O capítulo apresenta o banco de dados GEO (*Gene Expression Omnibus*), um repositório mundial que armazena e disponibiliza publicamente grandes quantidades de dados de expressão de gênica de câncer proveniente das amostras de pacientes. O capítulo mostra a importância de iniciativas como a 'Bioinformatics Virtual Coordination Network' e a necessidade de usar *software* de Bioinformática para poder pesquisar esses dados. Descreve especificamente a ferramenta de busca geoCancerPrognosticDatasetsRetriever, que recupera conjuntos de dados do GEO que contém informações sobre o prognóstico de um câncer específico. Assim, o capítulo apresenta uma área da Bioinformática que permite a criação de uma nova ferramenta computacional dedicada à análise de dados relacionados ao prognóstico de alguma categoria de câncer.

2.1 Bioinformática

A área de Bioinformática tem como principal objetivo criar e implementar métodos e ferramentas computacionais para o armazenamento, organização, análise e interpretação de grandes quantidades de dados que foram gerados por sistemas naturais. Para atingirem todo o seu potencial, a Biologia Computacional e a Bioinformática mantêm laços com os conhecimentos sobre a vida. A Bioinformática usa os princípios do conhecimento e a Tecnologia da Informação (TI) para tornar os vastos, diversos e complexos dados sobre a vida mais acessíveis e úteis. A Biologia Computacional aborda questões teóricas e experimentais usando métodos computacionais (NIH, 2000). Apesar das diferenças entre a Bioinformática e a Biologia Computacional, há muita interligação e esforço na interface (NIH, 2000; BIOINFO, 2022).

O National Institute of Health (NIH), através do Biomedical Information Science and Technology Initiative Consortium (BISTIC) chegou a um acordo sobre as delineações subsequentes a respeito da Bioinformática e da Biologia Computacional, considerando que nenhuma descrição poderia excluir totalmente a interseção com outras condicionantes ou evitar variações na interpretação por parte de diferentes individualidades e associações (NIH, 2000).

A investigação, criação ou aplicação de métodos e ferramentas computacionais para ampliar a utilização de dados biológicos, fisiológicos, psicológicos ou comportamentais, incluindo os que se destinam a recolher, arquivar, armazenar, catalogar, analisar ou visualizar dados semelhantes, são as principais funções da Bioinformática (NIH, 2000; INCA,2022). A Biologia Computacional abrange por sua vez, o desenvolvimento e utilização dos dados, modelos computacionais para o estudo de sistemas naturais, comportamentais e sociais (NIH, 2000).

Como resultado do aumento na quantidade dos dados, muitos dos desafios da Biologia tornaram-se, de fato, desafios da Computação. Esta abordagem é ideal para investigar as dinâmicas complexas da natureza porque os computadores podem tratar grandes quantidades de dados eficientemente. A Bioinformática, pode ser definida como a aplicação de técnicas computacionais para a compreensão e a organização de dados que estão relacionados a macromoléculas naturais. A natureza da biologia como TI explica principalmente esta associação entre os dois campos (ABRIATA *et al*., 2020; YAO *et al*., 2020; THE ECONOMIST, 1999). Ela permite a simulação de processos e estruturas das moléculas, proporcionando percepções importantes. Além disso, a Bioinformática tem sido fundamental nos estudos evolutivos, facilitando as investigações sobre a molécula de DNA (PROSDOCIMI, 2007).

Nos anos 70, os pesquisadores Ben Hesper e Paulien Hogeweg começaram a usar o termo "Bioinformática" para a exploração que queriam fazer, definindo como "estudo do processamento informacional dos sistemas bióticos" (apesar da afirmação de várias fontes públicas de que a expressão foi utilizada pela primeira vez em artigos a partir do ano de 1978 (HOGEWEG, HESPER,1978; HOGEWEG, 1978). O termo Bioinformática tornou-se mais popular a partir do final da década de 1980, quando foi substancialmente utilizado para relacionar estilos computacionais para análise relativa de dados genômicos. (HOGEWEG, 2011).

É fundamental definir os fundamentos que apoiam o sucesso dos especialistas no domínio da Bioinformática e da Biologia Computacional à medida que este se desenvolve e cresce. Os rápidos avanços na TI e nas Ciências Biológicas, exigem uma melhoria constante dos programas de formação em Bioinformática, de modo a preservar a sua relevância. É fundamental identificar os fundamentos que permitem aos profissionais do domínio da Bioinformática e da Biologia Computacional ter êxito à medida que este se desenvolve e cresce (WELCH *et al*, 2012). Essas individualidades trabalham numa grande variedade de contextos, incluindo laboratórios de Bioinformática, laboratórios de exploração natural e médica,

associações de desenvolvimento de *software*, empresas de desenvolvimento de medicamentos e instrumentos, e instituições que dão educação, serviço e formação (WELCH *et al*., 2014).

Apesar da Bioinformática possuir raízes no estudo das sequências de nucleotídeos e proteínas, ela inclui abordagens destinadas a modalidades de acesso aos dados e visa compreender o modo como os sistemas naturais funcionam numa variedade de contextos. Os bioinformatas, como são chamados os profissionais da área de Bioinformática, precisam estar conscientes de diferentes campos científicos, como a Biologia tradicional e molecular, Genética, Matemática, Estatística e Computação. Essa competência multidisciplinar é fundamental no desempenho de suas funções (ARON *et al*., 2021).

2.1.1 Combinação de conhecimentos em Biologia, Computação e Estatística na Bioinformática

Juntamente com a produção de dados, não há dúvida de que o gerenciamento, armazenamento e, consequentemente, recuperação, análise e interpretação de dados estão no centro de qualquer projeto de pesquisa biológica (KAUPPINEN, ESPINDOLA, 2011; GARIJO *et al*, 2013). Por exemplo, o sistema operacional Debian Med oferece pacotes de *software* coesos e prontos para informática médica e Bioinformática (MÖLLER *et al*., 2010). Outro exemplo de ferramentas utilizadas em Bioinformática é o RStudio Cloud, que permite programação na linguagem R (RSTUDIO CLOUD, 2021). Outro exemplo, é o *software* o QIIME2, que realiza análises de amplicons do microbioma (BOLYEN *et al*., 2019).

Um dos desafios da Bioinformática é lidar com a diversidade e complexidade dos dados disponíveis, como por exemplo, em um estudo metagenômico, onde deve-se comparar comunidades microbianas, utilizando sequências de DNA na identificação de organismos e genes contidos em amostras, interpretando os resultados biologicamente. Então foi criada uma plataforma em uma rede cooperativa chamada de "Bioinformatics Virtual Coordination Network" para ajudar os pesquisadores de bioinformática na comunicação e na organização entre eles para facilitar o compartilhamento de informações e de recursos e além de permitir que trabalhem em conjunto para resolver os mais diversos problemas (TULLY *et al*., 2021).

Para superar o desafio de lidar com grandes volumes de dados disponíveis, empreendimentos educacionais semelhantes ao Bioinformatics Virtual Coordination Network (BVCN) foram criados para educar e aprender Bioinformática durante a epidemia. Em março de 2020, o líder da BVCN, Dr. Benjamin Tully postou um anúncio no Twitter para avaliar o interesse em tutoria e Bioinformática durante a epidemia. Em poucos dias, o anúncio atraiu o

interesse de mais de 50 professores de Biologia Computacional e várias centenas de participantes. Depois de uma reunião virtual introdutória entre preceptores e interessados foi criando o espaço de trabalho BVCN Slack (TECKCHANDANI, 2018). No BVCN Slack, são abordados temas tais como: Unix, programação R, amplicons, metagenômica, transcriptômica, genética populacional e genômica e programação Python (TULLY *et al*., 2021).

Na pesquisa biológica, as dificuldades associadas à integração de dados só se expandiram com o advento de tecnologias de alto rendimento (MA'AYAN *et al*., 2014; SALEEM *et al*., 2014; KADADI *et al*., 2014). Projetos que utilizam Next Generation Sequencing (NGS) enfrentam desafios associados a uma variedade de aspectos que esse tipo de dados, principalmente o grande volume de dados (WANDELT *et al*., 2012; NEKRUTENKO, TAYLOR, 2012).

O principal objetivo da integração de dados é extrair conhecimento biológico a partir de conjuntos de dados que não podem ser obtidos conjuntamente (MOSTAFAVI *et al*., 2008; DUTKOWSKI *et al*, 2013). Compreender o processo celular e as interações moleculares por meio da integração de redes moleculares tem sido apenas um dos desafios da integração de dados (PRUITT *et al*., 2012). Um bom exemplo da importância dessa integração está relacionado a um método de detecção de via ativa, que utiliza programação dinâmica. Um algoritmo de caminho mais curto é aplicado para detectar caminhos que melhor explicam os dados de expressão gênica. Em um estudo de caso de Jiang *et al*. (2020), foi utilizado um conjunto de dados de câncer de pulmão humano que incluía dados de expressão gênica, dados de mutação e dados de metilação. Esses dados foram integrados a partir de uma abordagem de programação dinâmica que utiliza um algoritmo de caminho mais curto para identificar os caminhos que melhor explicam os dados de expressão gênica. Os resultados do estudo mostram que uma abordagem integrada pode identificar novas informações sobre os genes e os mecanismos que estão relacionados ao câncer de pulmão. Por exemplo, a investigação identificou uma nova via molecular que está envolvida na progressão do câncer de pulmão. Essa via molecular não havia sido descoberta anteriormente.

2.1.2 Linguagem de Programação Python

A linguagem de programação Python é uma ferramenta importante para a pesquisa científica. Ela é amplamente utilizada em uma variedade de áreas, incluindo desenvolvimento web, análise de dados, inteligência artificial e computação científica. O Python é uma

linguagem de programação de alto nível, de propósito geral e de código aberto, e possui uma sintaxe clara e concisa (BEAZLEY, 2009).

A linguagem de programação Python é bem versátil e pode ser usada para uma variedade de trabalhos de pesquisa científica. Ela é frequentemente usada para o desenvolvimento de software, análise de dados, visualização de dados, modelagem e simulação (BEAZLEY, JONES, 2013).

O Python é uma ferramenta que oferece muitos recursos computacionais para a análise de dados biológicos. Ela é fácil de aprender e de usar, além de possui várias possibilidades de aplicações (FOUNDATION, 2023).

2.3 *Gene Expression Omnibus* (GEO) e o geoCancerPrognosticDatasetsRetriever

Desde o início do ano 2000 tem-se disponibilizado conjuntos de dados de expressão gênica a partir de *microarrays* em repositórios na *Web*, como por exemplo o GEO (*Gene Expression Omnibus*), que é gerido pelo *National Institutes of Health* (NIH) (EDGAR *et al*., 2002). Com o crescimento da Internet e a disponibilidade de recursos computacionais cada vez mais acessíveis, foi disponibilizado no GEO um grande número de conjuntos de dados da expressão gênica de pacientes que foram diagnosticados com câncer.

Consequentemente, procurar por conjuntos de dados específicos no GEO é como procurar uma agulha num imenso palheiro. Atualmente, a quantidade de dados armazenados no GEO é muito grande. Para sanar essa dificuldade, foram propostas ferramentas para otimizar buscas nesse repositório e para que se possa chegar ao conhecimento de forma mais rápida e mais eficiente. Uma ferramenta que se propõe ajudar nesse desafio, relacionadas ao prognóstico de câncer, disponibilizada pelo GEO (*Gene Expression Omnibus*), é a ferramenta de busca "geoCancerPrognosticDatasetsRetriever" (ALAMEER, CHICCO, 2022).

A ferramenta do GEO "geoCancerPrognosticDatasetsRetriever" resolve parcialmente esse problema, permitindo que os utilizadores recuperem facilmente conjuntos de dados de expressão gênica de pacientes com câncer que incluam informações sobre o prognóstico em poucos minutos e com requisitos mínimos de recursos informáticos. Além disso, essa ferramenta permite que o utilizador especifique em que plataforma(s) de *microarray* os conjuntos de dados devem se basear, o que é uma restrição necessária para a compatibilidade entre conjuntos de dados. (CHEN *et al*., 2011).

O único *software* necessário para executar essa ferramenta de bioinformática é a linguagem de programação Perl (versão 5.8.0 ou superior), que vem pré-instalado no macOS e

também em sistemas operacionais Linux. Nos sistemas operacionais Microsoft Windows, a ferramenta pode ser instalada e utilizada através da camada de compatibilidade do Windows Subsystem for Linux (WSL).

Para instalar a ferramenta geoCancerPrognosticDatasetsRetriever, o usuário deverá se certificar previamente se todas as dependências estão instaladas de acordo com seu sistema operacional. O processo de instalação da ferramenta é o mesmo tanto no Linux/macOS quanto no Windows com o WSL (Windows Subsystem for Linux). O usuário poderá usar o CPAN (Comprehensive Perl Archive Network) para instalar o geoCancerPrognosticDatasetsRetriever executando o seguinte comando no terminal:

```
cpanm App::geoCancerPrognosticDatasetsRetriever
```

Informações e orientações mais detalhadas sobre como instalar a ferramenta geoCancerPrognosticDatasetsRetriever podem ser encontradas consultando o 'README' no repositório Perl MetaCPAN em https://metacpan.org/release/AALAMEER/App-geoCancerPrognosticDatasetsRetriever-1.02/source/README.md.

O geoCancerPrognosticDatasetsRetriever utiliza as bibliotecas de sistema cURL e Perl HomeDir. Essa ferramenta pode ser facilmente utilizada em um único comando, que pode ser executado em qualquer console de shell ou terminal de qualquer computador Linux ou Mac com acesso à Internet:

```
geoCancerPrognosticDatasetsRetriever -d "CANCER_TYPE" -p "PLATFORM_CODES" -f "/FOLDER_NAME/"
```

O parâmetro "CANCER_TYPE" deve ter o nome do tipo de câncer que o usuário pretende pesquisar. Embora não seja possível afirmar o número exato de todos os tipos e subtipos de câncer que estão disponíveis no GEO, pode-se encontrar por exemplo o Câncer de Próstata (*Prostate Cancer*), o de Mama (*Breast Cancer*), o de Pulmão (*Lung Cancer*), o Colorretal (*Colorectal Cancer*), o de Bexiga (*Bladder Cancer*), o Melanoma (*Melanoma Cancer*), a Leucemia (*Leukemia Cancer*), o de Fígado (*Liver Cancer*), o de Pancreas (*Pancreatic Cancer*). Para saber se o tipo de interesse está disponível, você pode acessar o site do GEO https://www.ncbi.nlm.nih.gov/geo/ e usar a função de pesquisa para procurar pelo tipo específico de câncer.

O parâmetro “PLATFORM_CODES” contém os códigos de acesso GEO, separados por espaços, das plataformas dos conjuntos de dados que o utilizador pretende pesquisar.

E o argumento “/FOLDER_NAME/” indica o caminho para o diretório onde o *script* deve guardar os arquivos temporários e de saída. Por exemplo, se o utilizador quiser saber os códigos de acesso GEO de qualquer conjunto de dados de prognóstico de dados de expressão gênica de pacientes diagnosticados com câncer de bexiga (*bladder cancer*) para as plataformas *Affymetrix Human Genome* U133 HG-U133 Plus 2 (GPL570), *Human Genome* U133 HGU133A (GPL96) e *Human Genome* U133 HG-U133B (GPL97), pode digitar em um terminal Linux o seguinte comando:

geoCancerPrognosticDatasetsRetriever -d ”bladder cancer” -p ”GPL570 GPL97 GPL96” -f ”/bladder_files/”

No final da execução, o *script* imprime o resultado tanto no terminal quanto no arquivo de saída na pasta /bladder_files/. No caso do exemplo do “*bladder cancer*” o *script* imprime as linhas de saídas a seguir:

Total prognostic datasets found: 2
[1] GSE31684
[2] GSE5287

Como se pode observar, a ferramenta bioinformática “geoCancerPrognosticDatasetsRetriever” encontrou no banco de dados do GEO dois arquivos que contém dados relacionados ao prognóstico de câncer de bexiga (ALAMEER, CHICCO, 2022).

2.4 Câncer Colorretal e Câncer de Mama

Segundo o National Cancer Institute (2021), o câncer é uma doença que se caracteriza pelo crescimento descontrolado de células anormais. Existem mais de 100 tipos de câncer, cada um com as suas próprias características. Para exemplificar o funcionamento do programa, foram selecionados dados genômicos do câncer colorretal e do câncer de mama. Esses são dois tipos de câncer que estão entre os mais comuns no mundo, e existe uma grande quantidade de

informações sobre eles que estão disponíveis no banco de dados Gene Expression Omnibus (GEO).

O câncer colorretal é o terceiro tipo de câncer mais comum no mundo, com uma incidência global de 1,3 milhões de casos em 2022. A doença é mais comum em adultos com mais de 50 anos, com uma taxa de incidência de 25,6 casos por 100.000 pessoas. E os fatores de risco para o câncer colorretal incluem:
Fatores de risco modificáveis:

Idade
Obesidade
Tabagismo
Dieta pobre em fibras e rica em gorduras

Fatores de risco não modificáveis:

História familiar de câncer colorretal
Doença inflamatória intestinal.
Polipose adenomatosa familiar (WHO, 2023; SOLIMAN, *et al*., 2023).

O prognóstico do câncer colorretal depende do estágio da doença no momento do diagnóstico. Os casos diagnosticados em estágio inicial, as chances de sobrevivência são de 90% a 95%. Os casos diagnosticados em estágio avançado, as chances de sobrevivência são de 5% a 10% (WHO, 2023).

O câncer de mama é o tipo de câncer mais comum entre as mulheres, com uma incidência global de 2,2 milhões de casos em 2022. A doença é mais comum entre as mulheres com mais de 50 anos, com um índice de 30,6 casos a cada 100.000 mulheres. E os fatores de risco para o câncer de mama incluem:

Fatores de risco modificáveis:

Idade
Obesidade
Tabagismo

Consumo excessivo de álcool

Fatores de risco não modificáveis:

História familiar de câncer de mama
Mutações genéticas BRCA1 e BRCA2
Exposição à radiação ionizante (ALKAABBAN, FERGUSON, 2022; CLINIC, 2023).

O prognóstico do câncer de mama também depende do estágio da doença no momento do diagnóstico. Os casos diagnosticados em estágio inicial, as chances de sobrevivência são de 90% a 95%. Os casos diagnosticados em estágio avançado, as chances de sobrevivência são de 20% a 30% (WHO, 2023).

O câncer é uma doença complexa que causa graves consequências na vida das pessoas. A escolha do tipo de câncer para exemplificar a pesquisa poderia ser qualquer tipo de câncer disponível no banco de dados do GEO em arquivos com o formato '.soft'. O câncer colorretal e o câncer de mama são dois dos tipos que afetam um grande número de pessoas. São doenças com múltiplos fatores de risco. E são cânceres bastante estudados, por pesquisadores e cientistas da área molecular (WHO, 2023).

2.5 Justificativa

O câncer é uma doença complexa que afeta milhões de pessoas em todo o mundo. A pesquisa do câncer é necessária para desenvolver métodos que possam melhorar a sobrevivência e a qualidade de vida dos pacientes. E uma ferramenta importante para a pesquisa do câncer é o geoCancerPrognosticDatasetsRetriever, que permite aos pesquisadores acessar o banco de dados do GEO e fazer o download os conjuntos de dados que possuem informações relacionadas ao prognóstico de câncer. No entanto, essa ferramenta retorna uma grande quantidade de dados em cada arquivo recuperado, o que dificulta consideravelmente a análise.

Esse projeto de pesquisa visa desenvolver uma nova ferramenta que para analisar e extrair dados específicos dos arquivos obtidos pelo geoCancerPrognosticDatasetsRetriever, com o potencial de contribuir significativamente para a pesquisa do câncer. Essa nova ferramenta computacional irá facilitar a análise dos dados e permitirá aos pesquisadores obter percepções importantes sobre o prognóstico do câncer que está sendo estudado, Ele poderá

ajudar no desenvolvimento de novas terapias, tratamentos personalizados e outras formas para combater a doença.

3 ARTIGO

Submission date: XX/XX/2023
Resubmission date: XX/XX/2023
Camera ready submission: 31/10/2023

1st round notification: XX/XX/2023
2nd round notification: XX/XX/2023
Available online: XX/XX/2023
Publication date: XX/XX/2023

Section: regular article

Development of Bioinformatics software written in Python programming language to read and extract information from geoCancerPrognosticDatasetsRetriever files

Fabiano Borges Pereira[1]

[1] Programa de Pós-Graduação em Biotecnologia (PPGBIOTEC) - Universidade Federal de Ubelândia (UFU) - Patos de Minas - MG - Brasil

fabianoslb@outlook.com

Abstract. *With the constant increase in gene expression data, it is becoming increasingly necessary to create Bioinformatics tools to obtain information efficiently. In this context, UNITE-1 Understanding Novel Information Through Expression was created, a Python-based software to search the GEO (Gene Expression Omnibus) datasets and extract information related to cancer prognosis. UNITE-1 processes GEO soft files obtained through the geoCancerPrognosticDatasetsRetriever tool, using keywords to find the data of interest, and generates a new file with the information found. The results presented demonstrate the potential of UNITE-1 in the study of gene expression and the clinicopathological characteristics that are associated with cancer prognosis, highlighting its potential for the identification of new cancer biomarkers.*

Keywords*. Gene Expression, Cancer Prognosis, UNITE-1, Data Extraction, Bioinformatics.*

Resumo. *Com o constante aumento nos dados de expressão gênica, é cada vez mais necessário a criação de ferramentas de Bioinformática para alcançar informações com eficiência. Neste contexto foi criado o UNITE-1 (Understanding Novel Information Through Expression), um software baseado em Python para pesquisar os conjuntos de dados do GEO (Gene Expression Omnibus) e extrair informações relacionadas ao prognóstico de câncer. O UNITE-1 processa arquivos soft do GEO que são obtidos através da ferramenta geoCancerPrognosticDatasetsRetriever, utilizando paravras-chave para encontrar os dados de interesse, e gera um novo arquivo com as informações encontradas. Os resultados apresentados demonstram o potencial do UNITE-1 no estudo da expressão gênica e as características clinicopatológicas que estão associadas ao prognóstico de câncer, destacando seu potencial para a identificação de novos biomarcadores no câncer.*

Palavras-Chave. *Expressão Gênica, Prognóstico de Câncer, UNITE-1, Extração de dados, Bioinformática.*

Cite as: Pereira, F. B. (2023). Development of Bioinformatics software in Python programming language to read and extract information from geoCancerPrognosticDatasetsRetiever files. iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), XX(X), X:X-X:XX. DOI: XX.XXXX/isys.2023.XXXX

1. Introdução

A área da Bioinformática vem representando um avanço tecnológico acelerado nos últimos tempos, que por sua vez, impulsiona as pesquisas biomédicas, ao mesmo tempo que gera rapidamente muitos novos dados que são armazenados em banco de dados. Esse fato disponibiliza de forma acessível aos pesquisadores, uma grande quantidade de informações sobre as características biológicas. Esse aumento exacerbado do volume de dados gera questionamentos, como por exemplo: como podemos extrair e reunir de um imenso arquivo de texto, somente as informações que são relevantes para uma determinada pesquisa de forma rápida, eficiente e organizada gerando um novo arquivo personalizado?

Um banco de dados que podemos citar como exemplo e que possibilitou essa pesquisa, é o GEO (*Gene Expression Omnibus*), que é um grande banco de dados que hospeda uma diversificada variedade de conjuntos de dados sobre a expressão dos genes, abrangendo os mais variados contextos biológicos. O GEO hospeda uma grande variedade de dados sobre câncer, incluindo dados de expressão gênica, dados de DNA, dados de RNA e dados de proteínas. Esses dados podem ser usados para estudar os mecanismos biológicos do câncer, como identificar novos biomarcadores e prever o prognóstico. A quantidade e a complexidade desses conjuntos ou arquivos de dados fazem da análise manual um trabalho árduo e desgastante, além de ser um procedimento que exige muito tempo e energia do pesquisador para que seja totalmente executado.

Embora o GEO contenha uma grande quantidade de dados genômicos e transcriptômicos de milhares de experimentos, a sua organização e a sua acessibilidade frequentemente se tornam um obstáculo para os pesquisadores que procuram por subconjuntos específicos de informações. E para amenizar esse problema, o foi proposta a ferramenta geoCancerPrognosticDatasetsRetriever (ALAMEER, CHICCO, 2022), que busca no banco de dados do GEO os arquivos que contém as informações de expressão gênica prognóstica que estão relacionadas com algum tipo de câncer. Porém, mesmo a ferramenta encontrando e fornecendo apenas os arquivos que contém dados sobre o prognóstico, ainda sim a quantidade de informações em cada conjunto de dados é muito grande.

É dentro deste cenário problemático que surgiu a criação de um *software*[1] como uma solução, que depois de concluído foi batizado de UNITE-1 (*Understanding Novel Information Through Expression*). O UNITE-1 tem como objetivo principal simplificar a extração de informações específicas de arquivos do banco de dados do GEO, como uma solução prática e eficiente voltada para os pesquisadores que trabalham com a análise de dados da expressão gênica que estão diretamente relacionados com o prognóstico e diagnóstico de algum tipo de câncer.

Nesse contexto, a ferramenta UNITE-1 foi criada, com a colaboração do modelo de linguagem de inteligência artificial GPT-4 criado pela OpenAI, para servir aos pesquisadores que trabalham com os dados de expressão genética relacionados com algum tipo de câncer processando os arquivos fornecidos previamente pela ferramenta geoCancerPrognosticDatasetsRetriever. Um dos principais objetivos do programa UNITE-1 é fornecer uma solução para analisar conteúdos específicos que estão disponíveis nos arquivos do GEO.

À medida que a tecnologia avança, o volume e a complexidade desses dados aumentam consideravelmente, assemelhando-se a um quebra-cabeça composto por uma grande quantidade de peças. Organizar, integrar e extrair as informações necessárias desses conjuntos de dados para uma pesquisa fica cada vez mais difícil. O algoritmo UNITE-1 se destaca como uma ferramenta que transpõe esse tipo de obstáculo, processando os arquivos que são previamente disponibilizados pela ferramenta geoCancerPrognosticDatasetsRetriever.

O geoCancerPrognosticDatasetsRetriever é uma ferramenta bioinformática que foi desenvolvida por Abbas Alameer e Davide Chicco com o objetivo de identificar os conjuntos de dados de expressão gênica com informações de prognóstico de câncer, a partir do banco de dados GEO (*Gene Expression Omnibus*), baseado no tipo de câncer que foi especificado como parâmetro de entrada, advindos de experimentos de *microarray*.

A ferramenta geoCancerPrognosticDatasetsRetriever está implementada em linguagem de programação Perl e pode ser instalada em sistemas operacionais Linux, macOS e Windows. A função da ferramenta é procurar com rapidez por conjuntos de dados relevantes no GEO, com base nos parâmetros de pesquisa que são fornecidos pelo usuário (ALAMEER, CHICCO, 2022).

[1] Nesse artigo, a palavra '*software*' é utilizada como sinônimo para 'programa de computador'.

2. Linguagem de Programação Python

A linguagem de programação Python é uma ferramenta computacional muito útil para a pesquisa científica. Ela é amplamente utilizada e abrange várias áreas, incluindo desenvolvimento web, análise de dados, inteligência artificial e computação científica (BEAZLEY, 2009).

Disponibilizando uma biblioteca com mais de 130.000 módulos de terceiros para melhorar ainda mais as suas capacidades, Python é uma linguagem de programação fácil de usar, e que vem sendo amplamente aplicada para os mais variados fins como o desenvolvimento web, a análise de dados, a Inteligência Artificial e a computação científica. Com a sua sintaxe e a sua tipagem dinâmica, a linguagem de programação Python tem a característica de apresentar uma codificação legível e bem rápida. Além disso, seu interpretador é compatível com o código C e com o código C++, tornando a integração ideal (EKMEKCI *et al*., 2016; HILL, 2020).

Tanto o interpretador quanto as bibliotecas Python, são de código aberto (EDUCBA, 2023) e funcionam em sistemas operacionais como o Linux, o Windows, e o macOS (KINSTA, 2023). Um código Python pode ser inserido e utilizado em executáveis autônomos ou também pode ser executado diretamente no interpretador interativo. A linguagem de programação Python proporciona um rápido desenvolvimento e ainda elimina o ciclo de compilação-link-execução que é necessário para usar os códigos escritos em linguagens de programação de nível inferior (FOUNDATION, 2023).

A linguagem de programação Python apresenta alguns tipos de dados de alto nível, como por exemplo, listas flexíveis e dicionários para que seja de fácil manipulação e que se tenha uma conveniente representação dos dados. Em comparação com os programas que são escritos em outras linguagens, como a linguagem de programação C ou a linguagem de programação Java, os programas que são escritos em linguagem de programação Python normalmente são de 3 a 5 vezes mais curtos por causa da utilização de alguns tipos de dados de alto nível. Além disso, a linguagem Python já vem com vários módulos destinados a execuções comuns como as de Entrada/Saída de arquivos, que são chamadas de sistema e kits de ferramentas GUI, e ainda permite que o usuário divida os programas Python em módulos reutilizáveis. Uma das características do Python é suportar vários paradigmas de programação, que inclui a programação procedimental, a programação orientada a objetos e também estilos de programação funcional (LUTZ, 2013).

Python é uma linguagem de programação que apresenta características como o tratamento de excepções, módulos, classes, tipos de dados dinâmicos de alto nível e tipagem dinâmica (BEAZLEY, 2009). Isso torna essa linguagem adequada para uma série de cenários voltados para a resolução de problemas (BEAZLEY, JONES, 2013). A escrita em Python pode ser utilizada como uma linguagem de *script* incorporada ou como uma linguagem autônoma voltada para as melhorarias das aplicações C/C++ (SUMMERFIELD, 2009).

O interpretador do código Python se integra perfeitamente em ambientes de desenvolvimento como por exemplo o PyCharm (JETBRAINS, 2023), o Jupyter Notebook (PROJECT JUPYTER, 2023) e o Spyder (SPYDER, 2023). Esses ambientes fornecem as ferramentas para escrever, executar e depurar um código escrito em Python (VAN ROSSUM; DRAKE, 2011). A filosofia de *design* de sintaxe do Python, se concentra na melhoria da legibilidade do código, utilizando a indentação para poder separar os blocos e manter uma disposição mais organizada. Os programas em Python são fáceis de se manter, tanto para os programadores quanto para os utilizadores (LUTZ, 2013; FOUNDATION, 2023). O Python é uma linguagem versátil e que pode ser usada para uma variedade de trabalhos na pesquisa científica. Ela é frequentemente usada para o desenvolvimento de software, análise de dados, visualização de dados, modelagem e simulação (BEAZLEY, JONES, 2013).

2.2. Configuração do Ambiente de Desenvolvimento

O ambiente de desenvolvimento utilizado para esta pesquisa é composto por duas ferramentas: o Anaconda e o Jupyter Notebook. Para a concepção do UNITE-1, o ambiente de desenvolvimento foi configurado com a versão 2023.03 do Anaconda (ANACONDA, 2023) e a versão 6.6.0 do Jupyter Notebook (JUPYTER, 2023).

O Anaconda é uma distribuição da linguagem de programação Python que oferece uma interface de linha de comando para a instalação e o gerenciamento de pacotes de ferramentas computacionais. O Anaconda é altamente empregado para entender os dados biológicos, como por exemplo, para identificar padrões genéticos ou mapear genomas (ANACONDA, 2023).

O Jupyter Notebook é um aplicativo que permite a execução de códigos Python. O software Jupyter Notebook oferece um ambiente de desenvolvimento de software interativo, disponibilizando funcionalidades que permitem que os usuários possam gerar e também compartilhar em tempo real os documentos que combinam código, equações, visualizações e conteúdos narrativos. Ele ainda consegue lidar com diversas linguagens de programação, como por exemplo Python, R, Julia e Scala (JUPYTER, 2023). Na Figura 1 é apresentado um fluxograma para a configuração do ambiente de desenvolvimento.

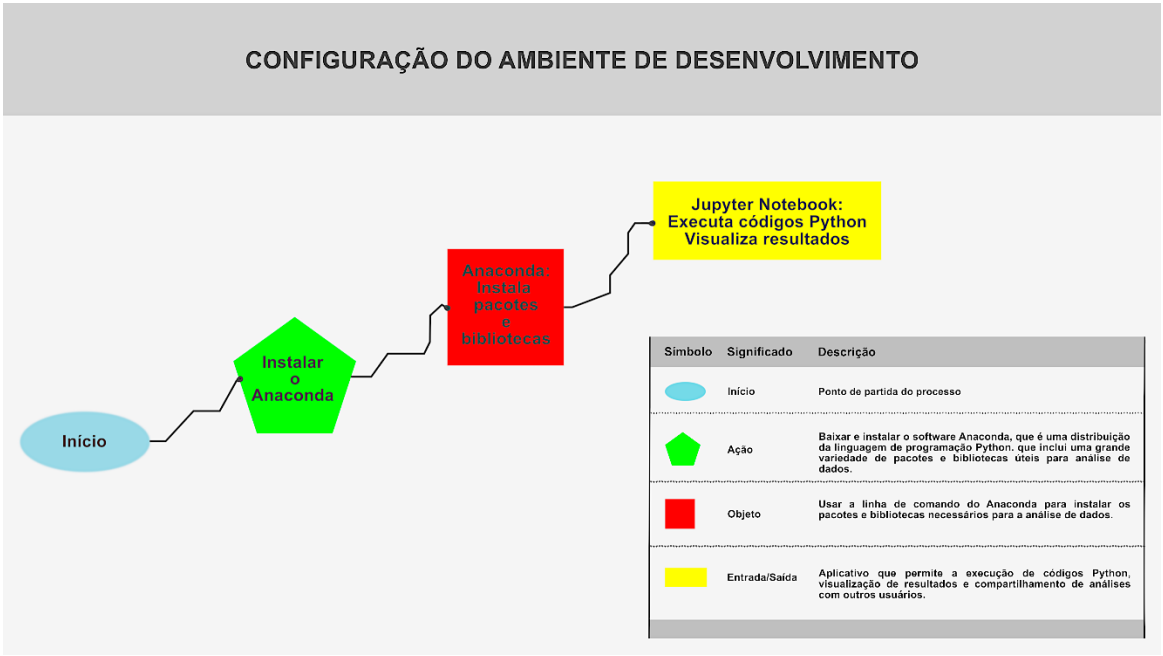

Figura 1: Fluxograma para a configuração do ambiente de desenvolvimento do UNITE-1.

Fonte: Imagem criada por Fabiano Borges Pereira.

3. Pacotes Python Utilizados no UNITE-1

O UNITE-1 é um programa de computador que permite a extração de informações prognósticas relacionadas ao tipo de câncer que o usuário está pesquisando. Para isso, o programa utiliza dois pacotes Python:

GEOparse
os

O pacote GEOparse permite a análise, a recuperação e a manipulação de dados do repositório GEO (Gene Expression Omnibus) (GUMIENNY, 2021). O GEO é um banco de dados público que contém informações de expressão gênica (NCBI, 2023).

O pacote os é um módulo da biblioteca padrão do Python (DOCUMENTATION, 2023) que disponibiliza algumas funções para interagir com o sistema operacional em que o Python está sendo executado. Ele é usado para a manipulação do sistema de arquivos e para a manipulação de caminhos (DOCUMENTATION, 2023; GEEKSFORGEEKS, 2022). Na Figura 2, é apresentado um fluxograma para os pacotes utilizados no UNITE-1.

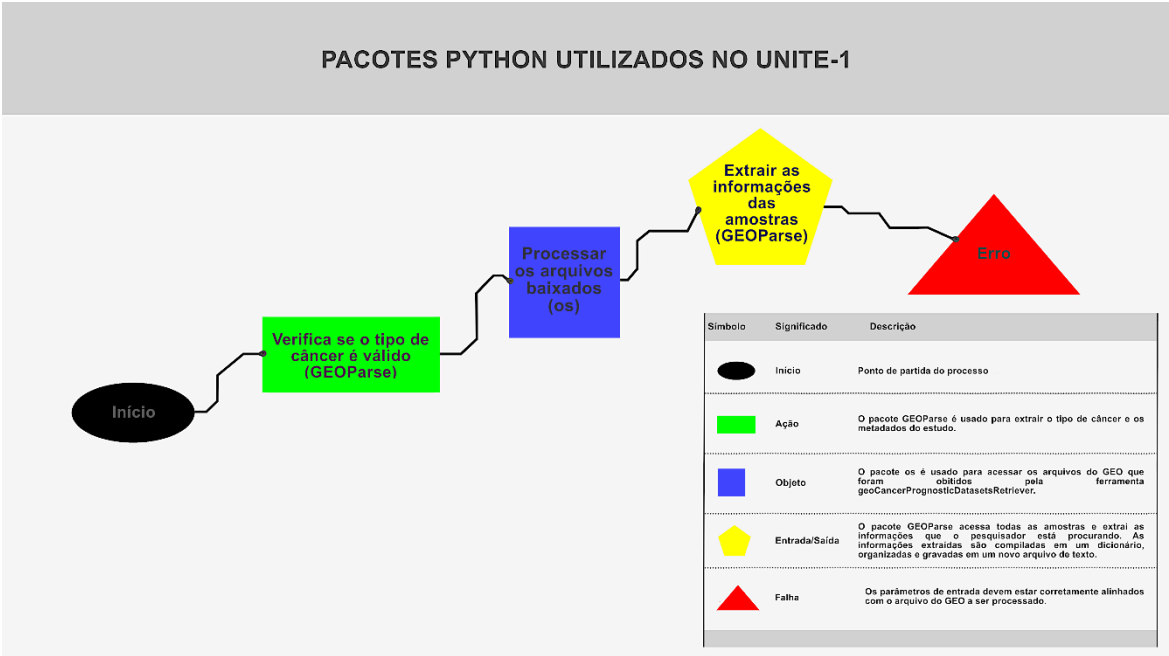

Figura 2: Fluxograma para a os pacotes Python utilizados no UNITE-1.

Fonte: Imagem criada por Fabiano Borges Pereira.

3.1. Discussão e a Aplicação dos Pacotes

O pacote 'GEOparse' e o pacote '*os*' estão presentes no algoritmo UNITE-1 e desempenham funções muito importantes que são indispensáveis para o bom funcionamento dessa ferramenta bioinformática. O pacote GEOparse interage com o arquivo do banco de dados GEO e faz o *download* dos conjuntos de dados de expressão gênica que foram previamente definidos. O pacote *os* por sua vez, vai construir os caminhos de arquivos para poder salvar os resultados e então lidar com as operações do sistema de arquivos.

Esses pacotes trabalhando juntos, recuperam com facilidade e eficiência os dados desejados contidos no arquivo do GEO, fazem um armazenamento bem estruturado dos resultados e faz uma manipulação flexível dos dados em formato txt. O pacote GEOparse vai acessar os conjuntos de dados iniciais, e o pacote *os* vai direcionar para onde os resultados deverão ser salvos. Cada pacote contribui com uma funcionalidade que é aproveitada para sustentar um algoritmo eficaz como o UNITE-1.

O uso desses pacotes no algoritmo UNITE-1, o torna capaz de coletar e processar automaticamente e com eficiência todos os dados de expressão gênica prognóstica do arquivo do GEO. Isso possibilita as meta-análises em larga escala dos dados prognósticos do câncer estudado. É um trabalho que, por meio de métodos manuais, não são viáveis. Os pacotes aumentam a precisão e padronizam o manuseio do conjunto de dados. Eles também melhoram a flexibilidade do UNITE-1 para que ele possa processar outros tipos de câncer por meio de pesquisas personalizadas.

A utilização desses pacotes reutilizáveis na programação traz benefícios, como por exemplo, a redução do tempo que é necessário para desenvolver um código e também a diminuição de erros. Isso possibilita realizar a criação de programas de computadores robustos mais rápida e segura, que poderão ser usados para extrair as informações a partir dos dados públicos gerados pela área biomédica.

Os pacotes GEOparse e o *os*, possuem funcionalidades distintas, uma é direcionada para acessar dados de um arquivo GEO e a outra para o gerenciamento de sistema de arquivos. Combinados de maneira adequada, os pacotes permitem a criação de um código mais automatizado e voltado para a mineração de conjuntos de dados que foram gerados pela expressão de genes prognósticos.

4. Utilização do UNITE-1 na Análise de Expressão Gênica Prognóstica de Câncer

O UNITE-1 (*Understanding Novel Information Through Expression*) é uma ferramenta bioinformática que foi criada com objetivo de buscar e extrair apenas as informações que o pesquisador precisa, gerando um novo arquivo personalizado e pronto para ser analisado, o que faz do UNITE-1 um recurso computacional nos conformes para os cientistas e os pesquisadores que estudam a expressão gênica do câncer.

O UNITE-1 é uma ferramenta bioinformática que foi desenvolvida para otimizar a pesquisa de fatores prognósticos em dados de expressão gênica provenientes de arquivos 'soft' do banco de dados do GEO, para diversos tipos de câncer. Para avaliar a capacidade e a eficiência do UNITE-1, foram processados com precisão os arquivos do banco de dados do GEO (*Gene Expression Omnibus*) que foram obtidos através da ferramenta geoCancerPrognosticDatasetsRetriever. Esses arquivos são referentes a dois tipos de câncer que tem uma alta relevância clínica, sendo eles o câncer colorretal com 10 arquivos recuperados e o câncer de mama com 33 arquivos recuperados, totalizando 43 conjuntos de dados.

4.1. geoCancerPrognosticDatasetsRetriever

O geoCancerPrognosticDatasetsRetriever é uma ferramenta computacional que foi desenvolvida com o objetivo principal de simplificar a identificação dos conjuntos de dados que contenha informações relacionadas ao prognóstico de vários tipos de câncer de plataformas de *microarray* que estão disponíveis no repositório do GEO (*Gene Expression Omnibus*). Com essa ferramenta não é necessário fazer demoradas pesquisas manuais no imenso banco de dados do GEO. Os usuários determinam o tipo de câncer e as plataformas que eles desejam e então a ferramenta vai recuperar os conjuntos de dados que são compatíveis.

4.1.1 Parâmetros de Entrada

Os parâmetros de entrada são as configurações fornececidas ao código antes de executá-lo, para que o seu funcionamento seja personalizado de forma que atenda às necessidades do usuário (DEVMEDIA, 2023).

Para encontrar os conjuntos de dados que tem informações relacionadas ao prognóstico de algum tipo de câncer usando o geoCancerPrognosticDatasetsRetriever, é preciso inserir algumas informações no *script* de comando para que os parâmetros de entrada e de saída sejam definidos. Para exemplificar usaremos o câncer de mama (*breast cancer*):

O tipo de câncer: -d "breast cancer"

Nesse parâmetro deverá ser definido o tipo de câncer do qual se deseja que a ferramenta pesquise no banco de dados, que nesse caso o geoCancerPrognosticDatasetsRetriever irá pesquisar o câncer de mama (*breast cancer*).

As plataformas de *microarray*: -p "GPL570 GPL97 GPL96"

Aqui é onde deverá ser especificado as plataformas de *microarray* que se deseja incluir na busca. São diferentes plataformas de *microarray* representadas pelos códigos "GPL570", "GPL97" e "GPL96" que serão usadas na pesquisa e coleta de dados genômicos. Nesse caso, a busca será por conjuntos de dados que sejam compatíveis com essas três plataformas.

E o diretório de saída: -f "/breast_files/"

O diretório de saída é o parâmetro que vai indicar o local aonde os resultados da busca serão armazenados. O caminho para o diretório onde os arquivos de saída serão salvos é representado por "/breast_files/", que pode ser acessado posteriormente.

```
$ geoCancerPrognosticDatasetsRetriever -d "breast cancer" -p "GPL570 GPL97 GPL96" -f "/breast_files/"
```

Depois de executar esse comando, o geoCancerPrognosticDatasetsRetriever vai procurar no banco de dados do GEOpor conjuntos de dados relacionados ao câncer de mama que sejam compatíveis com as plataformas

especificadas e que contenha informações sobre o prognóstico do câncer em questão. E é no diretório "/breast_files/" que os resultados da busca vão ser armazenados e ficarão disponíveis para o acesso posterior (Alameer & Chicco, 2022).

Na Figura 3, é mostrado o programa geoCancerPrognosticDatasetsRetriever em execução real no Windows Subsystem for Linux (WSL) através do Power Shell. A imagem exibe várias informações sobre o processo em andamento. Essas informações apresentadas na imagem fornecem dados sobre o funcionamento do programa, que inclui o *download* de arquivos, detalhes sobre as plataformas de *microarrays* e sobre as amostras que contém dados relacionados ao prognóstico do câncer que está sendo pesquisado, além da formatação e da análise dos dados.

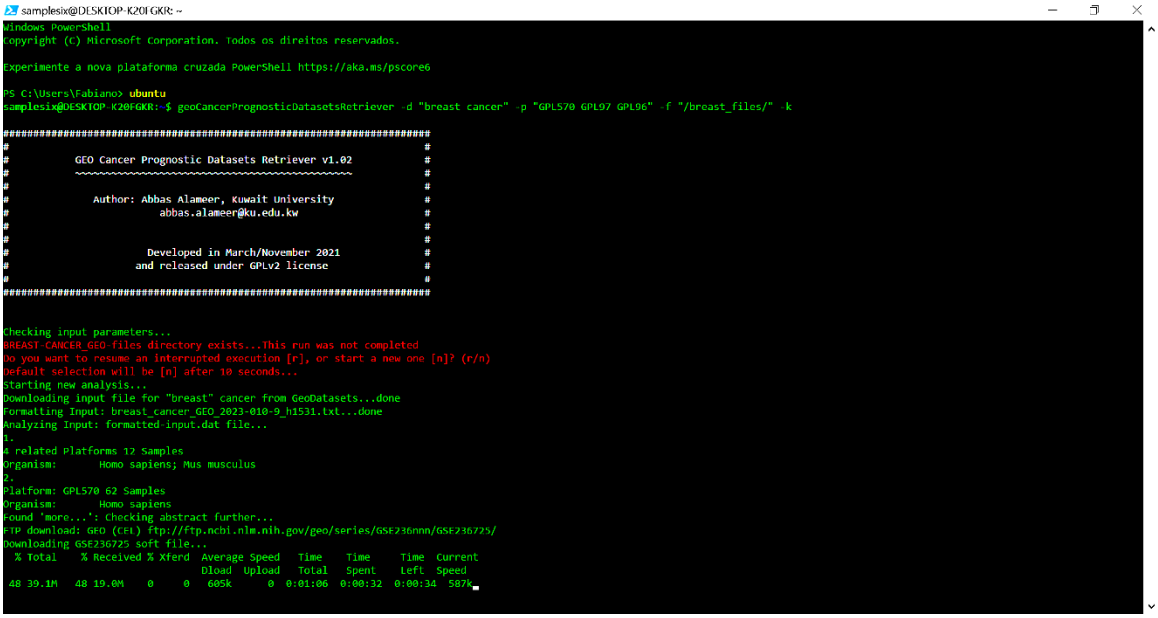

Figura 3: O programa geoCancerPrognosticDatasetsRetriever em execução no WSL pelo Power Shell.

Fonte: Captura de tela do geoCancerPrognosticDatasetsRetriever.

4.1.2 Resultados Gerados

A análise do geoCancerPrognosticDatasetsRetriever resultou na identificação de 33 conjuntos de dados relacionados ao câncer de mama que contém informações sobre o prognóstico do mesmo. Cada arquivo possui informações de expressão gênica e também informações clínicas, o que totaliza em 6,82 Gb de texto.

Para ter uma noção melhor do tamanho que essa quantidade de dados representa, imagine um livro de 600 páginas, com 30 linhas por página e 50 caracteres por linha. É um livro com aproximadamente 900.000 caracteres. Assim, considerando que normalmente cada caractere ocupa 1 byte, 6,82 gigabytes de texto correspondem aproximadamente à 7.578 livros de 600 páginas cada.

A Figura 4 mostra como a ferramenta entrega o resultado da pesquisa pelo câncer de interesse no GEO, em formato de lista. As informações apresentadas no resultado da execução da ferramenta geoCancerPrognosticDatasetsRetriever incluem dados sobre a conclusão da análise, o comando de entrada para executar o programa, a localização do arquivo com os resultados e a lista numerada dos arquivos que contém dados prognósticos que foram encontrados no banco de dados do GEO (*Gene Expression Omnibus*).

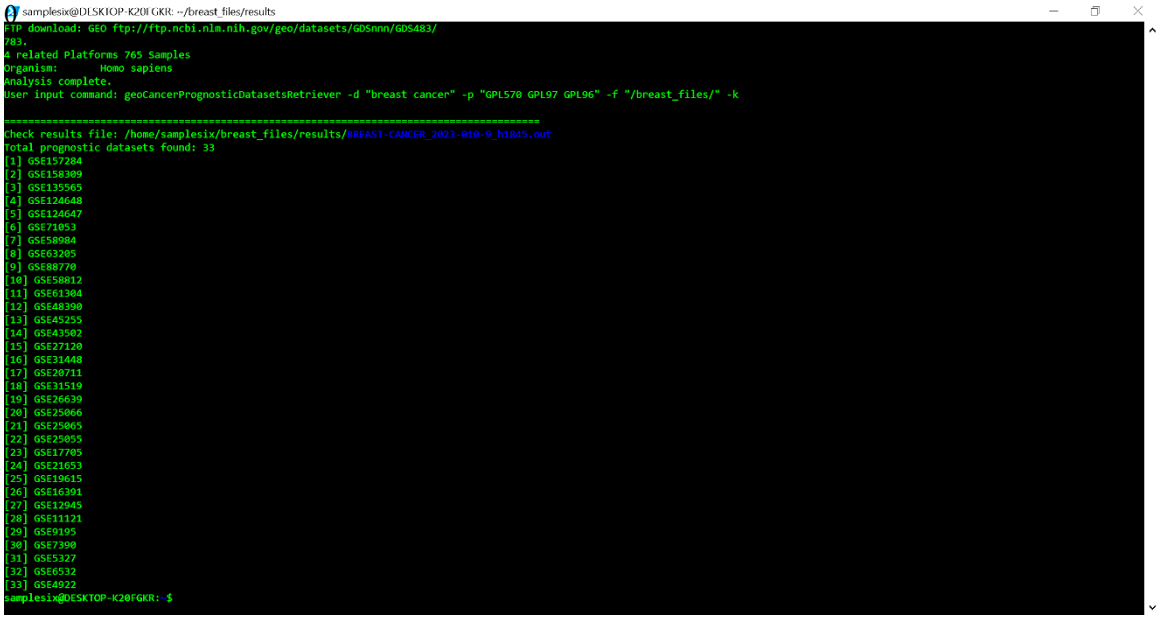

Figura 4: Resultado da pesquisa realizada pela ferramenta geoCancerPrognosticDatasetsRetriever.

Fonte: Captura de tela dos resultados obtidos.

Com a lista de códigos de acesso fornecida pelo geoCancerPrognosticDatasetsRetriever em mãos, o usuário pode acessar a página do GEO (https://www.ncbi.nlm.nih.gov/geo/) e pesquisar por cada código individualmente. O usuário pode inserir o código de acesso na barra de pesquisa na página inicial do GEO, que será direcionado para a página de registro do conjunto de dados referente ao código pesquisado. Nessa página, localize o link do arquivo com extensão SOFT que poderá ser identificado na seção "*Download family*" (Baixe a família). Quando o usuário clicar no link 'SOFT formatted family file(s)', ele será redirecionado automaticamente para a página em que o arquivo foi indexado. Assim, basta clicar no link que tem o código de acesso, por exemplo 'GSE157284_family.soft.gz', e o *download* automático começará. O arquivo SOFT é compactado e será salvo no computador do usuário. Logo após a conclusão do *download*, o usuário poderá utilizar um programa de computador como o 7-Zip ou o WinZip para que o arquivo .soft bruto seja descompactado.

A disponibilidade desses 33 conjuntos de dados fornece dados importantes para o avanço da pesquisa sobre os fatores de prognóstico e de biomarcadores para o câncer de mama. Porém, levando em consideração as limitações como a qualidade dos dados, a heterogeneidade da amostra e as tendências, uma boa e cuidadosa análise de cada conjunto de dados exigirá trabalho.

4.2 Parâmetros de Entrada do UNITE-1

A ferramenta UNITE-1 requer a configuração dos seguintes parâmetros, sendo um opcional e 8 obrigatórios para execução da análise e extração personalizada dos conjuntos de dados do GEO:

*header – possibilita inserir um cabeçalho personalizado.
"IDENTIFICAÇÃO DO ESTUDO"

O cabeçalho é uma identificação opcional que pode ser adicionada ao arquivo de saída do UNITE-1. O cabeçalho pode ser usado para fornecer informações sobre o usuário, o trabalho ou o resultado da análise.

*relevant_keywords – palavras-chave sobre características de interesse. Exemplo:

["prognosis",
"diagnosis",
"survival",
"treatment response",
"etc..."]

As palavras-chave são usadas para identificar as características de interesse no conjunto de dados do GEO. O UNITE-1 usa as palavras-chave para filtrar as características do conjunto de dados e extrair apenas as informações relevantes.

*extract_info – função que se baseia na lista de palavras-chave para encontrar e extrair as informações de interesse.
(tipo-câncer_cancer_keywords)

*relevant_title_keywords = (tipo-câncer_cancer_keywords)

*relevant_design_keywords = (tipo-câncer_cancer_keywords)

O tipo de câncer de interesse deverá ser definido. Nessa parte, o usuário coloca as palavras-chave que especifica o câncer que está sendo estudado, para que o título e os metadados de design possam ser verificados no conjunto de dados do GEO.

*filepath – caminho do arquivo do GEO a ser processado
os.path.join("C:", "arquivo.soft")

O local do arquivo é o caminho para o arquivo do GEO que será analisado. O UNITE-1 precisa ter acesso ao arquivo do GEO para extrair informações de interesse.

gse – variável que armazena o conjunto de dados carregado.

tipo-câncer_cancer_keywords – palavras-chave específicas ao tipo de câncer. Exemplo:

["cancer", "breast", "tumor", "malignant", "etc..."]

info = (tipo-câncer_cancer_keywords)

As palavras-chave para extração de dados são usadas para especificar as informações que o usuário deseja extrair. O UNITE-1 usa essas palavras-chave para encontrar as informações específicas que o usuário precisa. O tipo de câncer é usado para restringir a busca por conjuntos de dados relevantes.

*output_filepath – caminho para salvar arquivo de saída
os.path.join("C:", "resultados.txt")

O usuário cria um nome de identificação do novo arquivo e determina o caminho onde ele deseja salvar as informações que foram extraídas.

O correto preenchimento desses parâmetros obrigatórios, alinhados com o câncer estudado em cada caso, é vital para a personalização e o funcionamento correto do UNITE-1 na extração das informações de prognóstico presentes no arquivo do GEO que será processado. Por exemplo, se o usuário estiver pesquisando o câncer de mama, os parâmetros devem ser informados como ‘breast_cancer’ e as palavras-chave devem ser definidas em inglês e de acordo com o tipo ‘mama’, bem como o caminho que o arquivo do GEO está, e o caminho onde o novo arquivo deve ser salvo. O usuário pode executar o código no Jupyter Notebook.

Durante a execução, é mostrado cada etapa do processamento no arquivo do GEO. No canto superior esquerdo pode ser observado a data e hora do processamento. Logo depois, é apresentada a mensagem "INFO GEOparse Parsing C:\Users\Fabiano\Desktop\Breast\GSE157284_family.soft", que indica qual arquivo está sendo processado. Em seguida, são informados os registros de depuração (DEBUG) do GEOparse, que é a biblioteca utilizada para o programa poder acessar os dados do GEO. E na sequência é forncido informações

sobre o banco de dados utilizado (GeoMiame), a série (GSE157284) e a plataforma de microarray (GPL570) respectivamente. Na parte inferior da captura de tela, pode ser observado uma série de registros de depuração (DEBUG) relacionados às amostras (SAMPLE) que estão presentes no arquivo em processamento. Cada linha representa uma amostra individual de um paciente, que é identificada por um código (GSM) seguido de um número. A lista de amostras exibidas no resultado inclui as identificações GSM4760766, GSM4760767, GSM4760768, até a amostra GSM 411289. Totalizando 55 amostras de pacientes com câncer de mama encontradas no conjunto de dados GSE16391.

Em um notebook pessoal Aspire 5, 11th Gen Intel(R), processador Core(TM) i3-1115G4, 3.00GHz, em um sistema operacional Windows 10 Pro, versão 22H2, sem a necessidade de conexão à internet, a execução desse exemplo de conjunto de dados do câncer de mama leva em torno de 6 segundos para ser totalmente processado.

4.3. Resultados gerados pelo UNITE-1

O UNITE-1 processou com sucesso os arquivos do banco de dados do GEO (*Gene Expression Omnibus*) que foram obtidos através do *software* geoCancerPrognosticDatasetsRetriever. Durante o processamento, foram encontrados vários marcadores prognósticos, demonstrando que o UNITE-1 faz a análise da expressão gênica prognóstica com eficiência. Esses resultados reforçam o potencial apresentado pelo UNITE-1 em contribuir para o progresso das pesquisas no campo da oncologia.

Os resultados gerados pelo *software* UNITE-1 serão apresentados em duas partes. A primeira parte contém informações sobre o conjunto de dados, e a segunda parte contém as informações prognósticas relacionadas à analise individual de cada amostra.

A primeira parte dos resultados fornece uma visão geral do conjunto de dados GSE16391, que foi escolhido como exemplo para ilustrar os resultados das operações do UNITE-1 nos arquivos do GEO, e traz as seguintes informações:

O arquivo é intitulado de "GGI: um potencial preditor de recidiva para pacientes com câncer de mama tratados com terapia endócrina no ensaio BIG 1-98". Ele foi submetido em 02 de junho de 2009 e teve sua última atualização em 06 de junho de 2022. A plataforma de *microarray* foi a GPL570. O responsável pelo contato é o Benjamin Haibe-Kains, o departamento relacionado é o Princess Margaret Research do Instituto Princess Margaret Cancer Centre. E as palavras-chave que foram encontras no título e no *design* do estudo são 'câncer' e 'mama', que mostra a importância da especificidade para essa área de pesquisa.

A saída textual do arquivo que foi gerado pelo sistema UNITE-1 contendo a primeira parte dos resultados obtidos, pode ser visualizada na Figura 5.

```
GSE_Info:
        Accession: GSE16391
        Title: GGI: a potential predictor of relapse for endocrine-treated breast cancer patients in the BIG 1-98 trial
        Submission date: Jun 02 2009
        Last update date: Jun 06 2022
        Type: Expression profiling by array
        Overall design: Test whether the Gene expression Grade Index (GGI) is a potential predictor of relapse for
endocrine-treated breast cancer patients in the BIG 1-98 trial.
        Platform: GPL570
        Contact: Benjamin,,Haibe-Kains
        Department: Princess Margaret Research
        Institute: Princess Margaret Cancer Centre
        Relevant title keywords: ['cancer', 'breast']
        Relevant design keywords: ['cancer', 'breast']
```

Figura 5: Primeira parte dos resultados provenientes do *software* UNITE-1.

Fonte: Captura de tela da primeira parte do resultado.

A segunda parte dos resultados inclui os dados que foram encontrados nas amostras de cada paciente, fornecendo informações detalhadas sobre as características e os indicadores de prognóstico encontrados em cada amostra. Uma amostra específica foi selecionada para servir de modelo ilustrativo:

Amostra: GSM411235

Nome da amostra: BIG1_98_X1
Título: BIG1_98_X1
Fonte: BIG1_98_X1

Características da Amostra:

Nome da amostra: BIG1_98_X1
Controle de caso 0- caso, 1 - controle: 1
cluster.id: YT3
ggi: 577.55
Idade: 54
Status pós-menopausa: 1
Receptor de estrógeno e receptor de progesterona: 1
Quimioterapia neoadjuvante ajustada recebida: 0
Grau: 2
Terapia local: 1
Linfonodo: 1
Tratamento: 1
Tempo livre de recorrência após o tratamento (t rfs meses): 25.265
Status HER2: 0
Tamanho: 2
Tecido: Tumor primário de mama

Característica Relevante:

Grau: 2

Molécula: RNA total

No estudo, a amostra GSM411235 com o nome de BIG1_98_X1 foi classificada como controle (1). Especificamente, essa amostra pertence ao cluster YT3 e demonstra um indice de ggi no valor de 577.55. A paciente tem 54 anos de idade, está na fase pós-menopausa (status pós-menopausa: 1) e tem receptores positivos para estrógeno e progesterona (1). Ela não recebeu a quimioterapia neoadjuvante ajustada (0). O grau do tumor foi classificado como 2, isso significa que é uma extensão moderada da diferenciação celular. Foi administrada terapia local (1) e contatou-se que havia o envolvimento de linfonodo (1). A paciente foi submetida a tratamento (1), e o procedimento resultou em um período livre da doença de 25,265 meses. O status HER2 foi classificado como negativo (0). Foi avaliado o tamanho do tumor como 2, e o tecido examinado foi identificado como um tumor de mama primário. Uma apresentação do arquivo textual de saída produzido pelo *software* UNITE-1 pode ser vista na Figura 6, que exibe a amostra que foi selecionada para servir como exemplo ilustrativo.

```
Samples:
        Sample name: GSM411235
        Title: BIG1_98_X1
        Source name: BIG1_98_X1
        Characteristics:
                - sample name: BIG1_98_X1
                - case control 0- case, 1 - control: 1
                - cluster.id: YT3
                - ggi: 577.55
                - age: 54
                - post menopausal status: 1
                - er pgr: 1
                - adj neoadj chemotherapy received: 0
                - grade: 2
                - local therapy: 1
                - node: 1
                - treatment: 1
                - t rfs months: 25.265
                - her2 status: 0
                - size: 2
                - tissue: primary breast tumor
        Relevant characteristics:
                - grade: 2
        Molecule: total RNA
        Description: samplename: unique (anonymous) id for the patient
        Protocol: N/A

        Sample name: GSM411236
        Title: BIG1_98_X2
        Source name: BIG1_98_X2
        Characteristics:
                - sample name: BIG1_98_X2
                - case control 0- case, 1 - control: 0
                - cluster.id: NL1
                - ggi: 842.26
                - age: 65
```

Figura 6: Segunda parte dos resultados provenientes do *software* UNITE-1.

Fonte: Captura de tela da segunda parte do resultado.

4.4. Discussão

O câncer é uma enfermidade grave que tem como uma de suas características principais o crescimento exacerbado de células e pode se expandir e se desenvolver várias partes do corpo (CANCER.NET, 2019; NCI, 2021; WHO, 2023). As mutações genéticas que podem acontecer nas células e que vão afetar diretamente o controle do crescimento celular, geralmente são base para a instalação dessa doença em um indivíduo (NCI, 2021; COOPER, 2000).

São conhecidos mais de cem diferentes tipos de câncer, cada um com as suas próprias variações de características e de respostas aos tratamentos (COOPER, 2000). O processo conhecido como metástase, resulta em novos tumores, e ocorre devido a capacidade que essa doença tem de invadir e se desenvolver nos tecidos vizinhos e até mesmo em locais mais distantes do ponto de origem (NCI, 2021; CANCER.NET, 2019).

Para diagnosticar e definir o melhor tratamento possível, procedimentos como os exames físicos, a biópsia, o exame histopatológico, os exames de imagem, e os testes moleculares, são amplamente utilizados (CLINIC, 2022). Algumas opções mais comuns para tratar o câncer pode incluir a quimioterapia, a radioterapia, a cirurgia, a imunoterapia, a terapia direcionada e os medicamentos. (COOPER, 2000; CANCER.NET, 2019; WHO, 2023).

O grau do tumor é um importante indicador para estimar as probabilidades de progressão do câncer, além de ajudar a determinar o prognóstico dos pacientes (TANEJA, *et al*, 2010; LI, *et al*, 2023).

As informações obtidas sobre cada paciente, fornecem uma compreensão mais abrangente sobre os marcadores prognósticos encontrados em cada amostra, o que possibilita o desenvolvimento de estratégias de combate à doença.

O resultado apresentado pelo UNITE-1, com as informações desejadas em detalhes, ajuda a entender os padrões da expressão gênica e as características clínicas relacionadas a vários tipos de câncer, ficando evidente que o acesso rápido à dados específicos é muito importante e faz uma grande diferença em uma pesquisa.

Um estudo realizado por Cartwright *et al*. em 2014 mostrou que, para uma pesquisa sobre o prognóstico de câncer, é fundamental incluir algumas informações mais detalhadas sobre os pacientes, como por exemplo, as informações vitais, a idade e o estágio da doença. Com essas informações em mãos, os pesquisadores podem correlacionar a expressão gênica do câncer com as características clínicas dos pacientes e identificar os elementos que tem o potencial de afetar o prognóstico (NARRANDES, XU, 2018).

A capacidade de identificar as palavras-chave que foram definidas previamente tanto nos títulos quanto nas descrições dos conjuntos de dados, economiza bastante tempo e ajuda os pesquisadores a se concentrarem nos conjuntos de dados que são mais promissores bem como a selecionar informações que atendam diretamente aos objetivos de pesquisa.

Os dados que são coletados pelo UNITE-1 podem ser usados, por exemplo, para criar métodos novos e mais precisos para diagnosticar o câncer. De acordo com o National Cancer Institute (2021), a identificação de marcadores que podem prever o progresso do câncer vai depender da correlação entre as características clínicas e a expressão gênica do paciente.

O UNITE-1 permite que os usuários realizem as adaptações necessárias para atender as exigências de uma pesquisa, e isso em um campo com tantas variações quanto o do câncer, é uma característica que pode ser muito bem aproveitada, pois segundo os estudos de Louis e seus colaboradores em 2021, diferentes tipos de tumores podem exigir abordagens diferentes.

O estudo de NARRANDES e XU no ano de 2018, mostra que a extração eficiente de informações específicas da expressão gênica, também podem ser usadas para criar terapias personalizadas, uma vez que a compreensão da expressão gênica individualizada permite que os tratamentos sejam adaptados de acordo com o perfil de cada paciente.

Assim, o código UNITE-1 mostra que é de fato uma ferramenta computacional com potencial para ajudar na pesquisa de prognóstico do câncer, podendo ser usado para auxiliar o desenvolvimento de novos métodos de diagnóstico mais eficazes e também para o desenvolvimento de terapias personalizadas. Além disso, a abordagem do UNITE-1 mostra outras possibilidades, como por exemplo, o código pode servir de base para a criação e desenvolvimento de novas ferramentas computacionais voltadas para outros tipos de pesquisa ligadas à expressão gênica com o objetivo de resolver os problemas e então contribuir para o avanço e a evolução contínua das pesquisas no campo molecular e no combate às doenças.

4.5. Logotipo do UNITE-1

O programa de computador desenvolvido nesse trabalho é simbolizado por uma imagem que consegue capturar a essência do UNITE-1, que também foi criada pelo autor desse artigo. A imagem consiste em desenhos de duas moléculas de DNA descompactadas, em posdição vertical, de tamanhos diferentes e dispostas lado a lado, na cor verde (#00AA00) e com o fundo escuro (#001F00), simulando a estética dos monitores de fósforo verde da década de 70. As letras na fonte ‘Courier New’, em preto absoluto (#000000).

No desenho, à esquerda, há uma molécula de DNA maior que representa o arquivo GEO que é obtido através da ferramenta de busca geoCancerPrognosticDatasetsRetriever. Esse é o arquivo bruto e contém uma enorme quantidade de dados genômicos que estão prontos para serem filtrados. Logo à direita, tem um desenho de uma molécula de DNA menor que representa o novo arquivo gerado pelo UNITE-1, contendo somente os dados de interesse, que neste caso são apenas as informações relacionadas com o prognóstico extraídos do arquivo GEO original. A imagem do logotipo do UNITE-1 pode ser vista na Figura 7.

A sigla UNITE (*Understanding Novel Information Through Expression*) aponta para a grande importância da união entre cientistas e pesquisadores de diferentes áreas para poder lidar cada vez melhor com as problemáticas complexas que surgem no campo da investigação molecular. Como é confirmado por Bertel *et al*. em um estudo de 2022, a união entre especialistas de diferentes áreas é considerada fundamental para poder responder as mais variadas perguntas, bem como resolver problemas complexos.

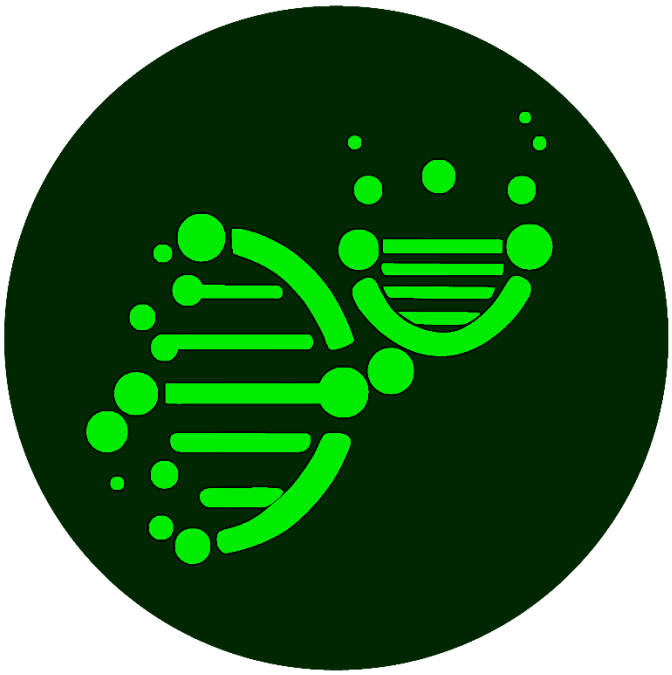

Figura 7: Logotipo do *software* UNITE-1.

Fonte: Logotipo também criado por Fabiano Borges Pereira.

5. Conclusões

O câncer é uma doença multifacetada caracterizada por apresentar um crescimento celular descontrolado, mutações genéticas, metástase e uma variedade de tratamentos. Essa complexidade traz dificuldades na pesquisa do câncer, mas também é uma área de grande potencial para o desenvolvimento de novos métodos de combate à doença.

O código UNITE-1 é uma ferramenta que pode ajudar a superar essas dificuldades. Ele é rápido na recuperação de dados específicos e apresenta uma busca eficiente das informações relacionados ao prognóstico do câncer nos conjuntos de dados do GEO, gerando assim um novo arquivo, permitindo que os pesquisadores obtenham as informações de interesse muitas vezes mais rápido que a busca manual.

A eficiência do código se destaca por processar o arquivo em questão de segundos, o que proporciona aos pesquisadores um acesso rápido às informações sobre o prognóstico do câncer que está sendo estudado. Além disso, o código não depende da internet para funcionar, basta ter disponível em seu computador o arquivo do GEO previamente baixado e descompactado.

É preciso destacar que o UNITE-1 pode analisar os arquivos do GEO (*Gene Expression Omnibus*) relacionados a qualquer forma de câncer, apesar de ter sido rigorosamente testado em conjuntos de dados de câncer colorretal e de mama. O potencial do UNITE-1 como uma ferramenta de análise da expressão gênica prognóstica em vários tipos de câncer é bastante aumentado por essa flexibilidade.

O UNITE-1 não se limita a uma única aplicação, mas sim a todos os tipos de câncer disponíveis no banco de dados do GEO (*Gene Expression Omnibus*) em arquivos '.soft'. Além disso, ele tem a capacidade de contribuir para o avanço do conhecimento e do tratamento do câncer, consolidando o UNITE-1 como uma ferramenta computacional para os pesquisadores que fazem a investigação oncológica.

Referências

ALAMEER, Abbas; CHICCO, Davide. geoCancerPrognosticDatasetsRetriever: a bioinformatics tool to easily identify cancer prognostic datasets on Gene Expression Omnibus (GEO). Bioinformatics, v. 38, n. 6, p. 1761–1763, mar. 2022. Disponível em: https://doi.org/10.1093/bioinformatics/btab852. Acesso em: 09 out. 2023.

ANACONDA. Anaconda Distribution. Free Download, 2023. Disponível em: https://www.anaconda.com/download. Acesso em: 25 set. 2023.

ANACONDA. HCC / packages / geoparse. Python library to access Gene Expression Omnibus Database (GEO), 2023. Disponível em: https://anaconda.org/hcc/geoparse. Acesso em: 1 out. 2023.

ANACONDA. The Minnesota Supercomputing Institute, 2023. Disponível em: https://www.msi.umn.edu/sw/anaconda. Acesso em: 26 set. 2023.

BEAZLEY, D. Python Essential Reference. 4. ed. Boston: Addison-Wesley Professional, 2009. ISBN-13: 978-0-672-32862-6. Disponível em: https://github.com/neocode/Easylabs/blob/master/Task/Python.Essential.Reference.4th.Edition.David.M.Beazley.2009.pdf. Acesso em: 26 set. 2023.

BERTEL, L. B.; WINTHER, M.; ROUTE, H. W.; KOLMOS, A. Framing and facilitating complex problem-solving competences in interdisciplinary megaprojects: An institutional strategy to educate for sustainable development. International Journal of Sustainability in Higher Education, 2022. V. 23, n. 5, p. 1173-1191. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/IJSHE-10-2020-0423/full/html. Acesso em: 20 out. 2023.

CANCER.NET. What is Cancer? 2019. Disponível em: https://www.cancer.net/navigating-cancer-care/cancer-basics/what-cancer. Acesso em: 06 out. 2023.

CARTWRIGHT, L. A.; DUMENCI, L.; SIMINOFF, L. A.; MATSUYAMA, R. K. Cancer Patients' Understanding of Prognostic Information. Journal of Cancer Education: The Official Journal of the American Association for Cancer Education, 2014. v. 29, n. 2, p. 311-317. Disponível em: https://doi.org/10.1007/s13187-013-0603-9. Acesso em: 19 out. 2023.

CLINIC, Mayo. Cancer. Overview, 2022. Disponível em: https://www.mayoclinic.org/diseases-conditions/cancer/symptoms-causes/syc-20370588. Acesso em: 06 out. 2023.

COOPER, G. M. The Cell: A Molecular Approach. 2ª ed. Sunderland (MA): Sinauer Associates, 2000. The Development and Causes of Cancer. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK9963/. Acesso em: 06 out. 2023.

DEVMEDIA. Principais conceitos da Programação Orientada a Objetos, 2023. Disponível em: https://www.devmedia.com.br/principais-conceitos-da-programacao-orientada-a-objetos/32285. Acesso em 26 nov. 2023.

DOCUMENTATION, Jupyter. Installing Jupyter Notebook, 2017. Disponível em: https://test-jupyter.readthedocs.io/en/latest/install.html. Acesso em: 27 set. 2023.

DOCUMENTATION, Python. os - Miscellaneous operating system interfaces, 2023. Disponível em: https://docs.python.org/3/library/os.html. Acesso em: 3 out. 2023.

DRISCOLL, M. Jupyter Notebook: An Introduction, 2019. Disponível em: https://realpython.com/jupyter-notebook-introduction/. Acesso em: 28 set. 2023.

EDUCBA. Why Python is Open Source Language? 2023. Disponível em: https://www.educba.com/python-is-open-source-language/. Acesso em: 19 de setembro de 2023.

EKMEKCI, B.; MCANANY, C. E.; MURA, C. An Introduction to Programming for Bioscientists: A Python-Based Primer. PLoS Comput Biol, v. 12, n. 6, e1004867, 2016. Disponível em: https://doi.org/10.1371/journal.pcbi.1004867. Acesso em: 19 de setembro de 2023.

FOUNDATION, Python Software. Documentation. Python Tutorial. 6. Modules, 2023. Disponível em: https://docs.python.org/pt-br/3/tutorial/modules.html. Acesso em: 1 out. 2023.

FOUNDATION, Python Software. The Python Tutorial, 2023. Disponível em: https://docs.python.org/3/tutorial/index.html. Acesso em: 18 set. 2023.

GEEKSFORGEEKS. How to install Jupyter Notebook on Windows?, 2021. Disponível em: https://www.geeksforgeeks.org/how-to-install-jupyter-notebook-in-windows/. Acesso em: 27 set. 2023.

GEEKSFORGEEKS. OS Module in Python with Examples, 2022. Disponível em: https://www.geeksforgeeks.org/os-module-python-examples/. Acesso em: 1 out. 2023.

GUMIENNY, R. GEOparse Documentation Release 1.2.0, 2019. [s.l: s.n.]. Disponível em: https://readthedocs.org/projects/geoparse/downloads/pdf/latest/. Acesso em: 30 set. 2023.

GUMIENNY, R. GEOparse. Python library to access Gene Expression Omnibus Database (GEO), 2021. Disponível em: https://pypi.org/project/GEOparse/. Acesso em 29 set. 2023.

GÜSE, J. How to set up Anaconda and Jupyter Notebook the right way, 2021. Disponível em: https://towardsdatascience.com/how-to-set-up-anaconda-and-jupyter-notebook-the-right-way-de3b7623ea4a. Acesso em: 27 set. 2023.

HILL, C. Learning Scientific Programming with Python. 2. ed. Cambridge: Cambridge University Press, 2020. Disponível em: https://doi.org/10.1017/9781108778039. Acesso em: 26 set. 2023.

JETBRAINS. PyCharm: The Python IDE for Professional Developers. Disponível em: https://www.jetbrains.com/pycharm/. Acesso em: 20 set. 2023.

JUPYTER. Installing Jupyter. Get up and running on your computer, 2023. Disponível em: https://jupyter.org/install. Acesso em: 27 set. 2023.

JUPYTER. Project Jupyter, 2023. Disponível em: https://jupyter.org/. Acesso em: 28 set. 2023.

KINSTA. How to Install Python on Windows, macOS, and Linux. 2023. Disponível em: https://kinsta.com/knowledgebase/install-python/. Acesso em: 19 de setembro de 2023.

LI, Y. H.; WANG, X. Y.; SHEN, J. W.; MA, L. L.; WANG, C. P.; HE, K.; LIU, D. S.; LI, Y. F. Clinical factors affecting the long-term survival of breast cancer patients. The Journal of international medical research, 51(3), 3000605231164004, 2023. Disponível em: https://doi.org/10.1177/03000605231164004. Acesso em: 28 nov. 2023.

LOUIS, D. N.; PERRY, A.; WESSELING, P.; BRAT, D. J.; CREE, I. A.; FIGARELLA-BRANGER, D.; HAWKINS, C.; NG, H. K.; PFISTER, S. M.; REIFENBERGER, G.; SOFFIETTI, R.; VON DEIMLING, A.; ELLISON, D. W. The 2021 WHO Classification of Tumors of the Central Nervous FBP System: a summary. Neuro-oncology, 2021. 23(8), 1231–1251. Disponível em: https://doi.org/10.1093/neuonc/noab106. Acesso em: 19 out. 2023.

LUTZ, M. Learning Python: Powerful Object-Oriented Programming. 5. ed. Sebastopol: O'Reilly Media, Inc., 2013. ISBN: 978-1-449-35573-9. Disponível em: https://github.com/Quyaz/books/blob/master/Learning%20Python%2C%205th%20Edition.pdf. Acesso em: 28 nov. 2023.

BEAZLEY, D.; JONES, B. K. Python Cookbook. 3. ed. Sebastopol: O'Reilly Media, Inc., 2013. ISBN: 978-1-449-34037-7. Disponível em: https://github.com/lpvcpp/learn_python/blob/master/D.%20Beazley%2C%20B.K.%20Jones%20-%20Python%20Cookbook%2C%203rd%20Edition.%202013.pdf. Acesso em 19 out. 2023.

NARRANDES, S.; XU, W. Gene Expression Detection Assay for Cancer Clinical Use. Journal of Cancer, 2018. 9(13), 2249-2265. Disponível em: https://doi.org/10.7150/jca.24744. Acesso em: 19 out. 2023.

NATIONAL CANCER INSTITUTE. Tumor Markers, 2021. Disponível em: https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-fact-sheet. Acesso em 19 out. 2023.

NATIONAL CANCER INSTITUTE. What is cancer? 2021. Disponível em: https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Acesso em: 06 out. 2023.

NCBI. Gene Expression Omnibus, 2023. Disponível em: https://www.ncbi.nlm.nih.gov/geo/. Acesso em 29 set. 2023.

PROJECT JUPYTER. Jupyter Notebook. Disponível em: https://jupyter.org/. Acesso em: 20 set. 2023.

PYTHON, Anaconda. Software. UNLV Information Technology, 2023. Disponível em: https://www.it.unlv.edu/software/anaconda-python. Acesso em: 26 set. 2023.

SPYDER. The Scientific Python Development Environment. Disponível em: https://www.spyder-ide.org/. Acesso em: 20 set. 2023.

SUMMERFIELD, M. Programming in Python 3: A Complete Introduction to the Python Language. 2. ed. Boston: Addison-Wesley Professional, 2009. ISBN 978-0-321-68056-3. Disponível em: https://github.com/huangblue/hello-world/blob/master/Programming%20in%20Python%203%20-%20A%20Complete%20Introduction%20to%20the%20Python%20Language%2C%20Second%20Edition%20(2010)%201.pdf. Acesso em: 26 set. 2023.

TAGLIAFERRI, L. Data Analysis and Visualization with pandas and Jupyter Notebook in Python 3. DigitalOcean, 2017. Disponível em: https://www.digitalocean.com/community/tutorials/data-analysis-and-visualization-with-pandas-and-jupyter-notebook-in-python-3. Acesso em: 28 set. 2023.

TANEJA, P.; MAGLIC, D.; KAI, F.; ZHU, S.; KENDIG, R. D.; FRY, E. A.; INOUE, K. Classical and Novel Prognostic Markers for Breast Cancer and their Clinical Significance. Clinical Medicine Insights. Oncology, v. 4, p. 15-34, 2010. Disponível em: https://doi.org/10.4137/cmo.s4773. Acesso em: 28 nov. 2023.

TECHWIKI.ENG.UA.EDU. Anaconda-Python, 2022. Disponível em: https://techwiki.eng.ua.edu/index.php/Anaconda-Python. Acesso em: 26 set. 2023.

VAN ROSSUM, G.; DRAKE, F. L. The Python Language Reference Manual. Bristol: Network Theory Ltd., 2011.

VERMA, I. Introduction to machine learning with Jupyter notebooks, 2021. Disponível em: https://developers.redhat.com/articles/2021/05/21/introduction-machine-learning-jupyter-notebooks. Acesso em: 28 set. 2023.

WICKRAMASINGHE, S. Jupyter Notebooks for Data Analytics: A Beginner's Guide, 2021. Disponível em: https://www.bmc.com/blogs/installing-jupyter-for-big-data-and-analytics/. Acesso em: 28 set. 2023.

WIKIPEDIA. Project Jupyter, 2023. Disponível em: https://en.wikipedia.org/wiki/Project_Jupyter . Acesso em: 28 set. 2023.

WORLD HEALTH ORGANIZATION. Cancer, 2023. Disponível em: https://www.who.int/health-topics/cancer#tab=tab_1. Acesso em: 06 out. 2023.

4 CONCLUSÕES

O desenvolvimento do UNITE-1, mostra que a utilização de recursos computacionais, pode auxiliar no desenvolvimento de trabalhos de pesquisa complexos e com grande volume de dados. O geoCancerPrognosticDatasetsRetriever fornece aos pesquisadores acesso direcionado a grandes conjuntos de dados de expressão gênica de câncer que contém informações relacionadas ao prognóstico da doença provenientes do banco de dados público Gene Expression Omnibus (GEO). No entanto, para extrair manualmente as informações de prognósticos nesses grandes conjuntos de dados do GEO exige muito tempo. O UNITE-1 é um software escrito em linguagem de programação Python personalizável desenvolvido para automatizar a extração de informações de prognóstico de qualquer tipo de câncer que estejam disponíveis em arquivos do GEO formatados em extensão '.soft'.

O algoritmo do UNITE-1 permite realizar pesquisas personalizadas e extrair rapidamente os dados de expressão gênica relacionados ao prognóstico do câncer nos arquivos do GEO, com base em uma lista de palavras-chave definidas pelo pesquisador. Os testes mostraram que o UNITE-1 pode processar os arquivos ‘.soft’ do GEO em questão de segundos, o que demonstra as suas vantagens de velocidade e de eficiência em relação à pesquisa manual. O código não possui dependência da Internet e funciona off-line com os arquivos '.soft' previamente baixados, otimizando ainda mais a facilidade de uso, além de diminuir consideravelmente a margem de erro.

Os conjuntos de dados do GEO que possuem informações relacionadas ao prognóstico do câncer podem conter vários gigabytes de resultados da expressão gênica, o que gera grandes dificuldades para a análise manual. O UNITE-1 permite que os pesquisadores possam encontrar as informações de interesse usando a lista de palavras-chave que foram previamente definas. Com o UNITE-1 os pesquisadores obtêm rapidamente as informações de que precisam, economizando consideravelmente o tempo do pesquisador. Além disso, o UNITE-1 pode ajudar a encontrar novos tratamentos ou personalizar o tratamento de acordo com o perfil gênico de cada paciente, ao invés de passar horas e horas pesquisando manualmente, os pesquisadores podem direcionar sua energia para aprofundar em suas análises e assim obter resultados melhores.

O sucesso desse projeto demonstra a capacidade que a linguagem de programação Python tem no desenvolvimento de algoritmos acessíveis voltados para fluxos de trabalho em bioinformática complexos e volumosos. Pesquisadores de várias áreas diferentes podem aproveitar a versatilidade da linguagem de programação Python para criar códigos

personalizados que reinventam positivamente as suas pesquisas, mesmo sem possuir um conhecimento avançado na área da programação.

Pode-se destacar também o potencial que o código tem de servir de base para a criação e o desenvolvimento de ferramentas ainda mais poderosas e direto ao ponto por parte dos pesquisadores que buscam responder os mais diversos questionamentos e resolver cada vez mais os problemas que surgem na área da bioinformática, ou até mesmo um ‘upgrade’ no UNITE-1.

A flexibilidade do UNITE-1 o torna amplamente aplicável à todos os tipos de câncer que estão disponíveis no GEO armazenados em arquivos ‘.soft’. O UNITE-1 se concentra rapidamente somente nos dados de prognóstico desejados, mostrando um potencial para acelerar descobertas e novos tratamentos. Esse trabalho apresentou o UNITE-1 como uma solução computacional para aproveitar efetivamente os dados públicos do GEO e avançar na pesquisa do prognóstico de câncer por meio da análise otimizada da expressão gênica.

5 PESPECTIVAS

O campo da Bioinformática é uma área de evolução acelerada, apresentando diversas oportunidades para pesquisa e para desenvolvimentos futuros. Embora o programa de computador desenvolvido tenha mostrado sua eficiência ao identificar os termos pré-definidos e extrair apenas as informações sobre o diagnóstico do câncer estudado, ainda há possibilidades de melhorias.

Um dos objetivos do UNITE-1 foi identificar as palavras-chave voltadas para o prognóstico de algum tipo de câncer. Além disso, ele objetivou poder ser possível sua personalização para poder lidar com outros tipos de informações sobre o câncer, ou para lidar com outros tipos de doenças. Adaptar o algoritmo para lidar com diversos tipos de dados, de outras fontes pode ser uma perspectiva interessante a ser considerada, além da possibilidade de se criar uma simplificação do programa, em vários aspectos, para que ele se torne bem mais acessível para usuários sem conhecimento avançado em programação, o que pode torná-lo mais acessível e mais utilizado pela comunidade científica.

Uma provável melhoria seria a implementação de técnicas de Aprendizado de Máquina, com o uso de Redes Neurais Artificiais por exemplo, para encontrar padrões nos dados gerados relacionados ao prognóstico de câncer, além poder ajudar na identificação de percepções importantes.

Uma outra perspectiva promissora é a união das ferramentas geoCancerPrognosticDatasetsRetriever (Perl) e UNITE-1 (Python) em um único programa que combine todas as funcionalidades das duas ferramentas em uma solução completa e integrada. Essa fusão simplifica o processo de busca e extração de dados, o que proporciona uma otimização ainda maior para a pesquisa de informações relacionadas ao prognóstico de câncer. A junção dessas duas ferramentas pode ser um princípio em potencial para revolucionar a forma como os pesquisadores buscam e exploram os conjuntos de dados do GEO. Isso proporcionaria avanços significativos na área da Bioinformática, contribuindo para o aumento da compreensão da expressão gênica, devido à possibilidade de adaptação do novo programa para pesquisas além dos prognósticos de câncer.

REFERÊNCIAS

ABRIATA, L. A.; LEPORE, R.; DAL PERARO, M. About the need to make computational models of biological macromolecules available and discoverable. Bioinformatics, v. 36, n. 9, p. 2952-2954, 2020. Disponível em: https://doi.org/10.1093/bioinformatics/btaa086. Acesso em: 06 ago. 2023.

ALAMEER, A.; CHICCO, D. Gene expression geoCancerPrognosticDatasetsRetriever: a bioinformatics tool to easily identify cancer prognostic datasets on Gene Expression Omnibus (GEO). Bioinformatics, v. 38, n. 6, p. 1761-1763, 2022. Disponível em: https://doi.org/10.1093/bioinformatics/btab852. Acesso em: 13 ago 2023.

ALKABBAN, F. M.; FERGUSON, T. Breast Cancer, 2022. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK482286/. Acesso em: 06 out. 2023.

ARON, S.; JONGENEEL CV, CHAUKE PA, CHAOUCH M, KUMUTHINI J, ZASS L, *et al*. Ten simple rules for developing bioinformatics capacity at an academic institution. PLoS Comput Biol, v. 17, n. 12, e1009592, 2021. Disponível em: https://doi.org/10.1371/journal.pcbi.1009592. Acesso em: 19 ago. 2023.

BALAJI, S. R.; GUPTA, K. K.; ANUSHA, P.; RAVEENA, P. Molecular Docking Studies of Wide Spectrum Targets in Staphylococcus aureus - An Aim towards Finding Potent Inhibitors. Adv Tech Biol Med, v. 2, p. 115, 2014. Disponível em: https://doi.org/10.4172/2379-1764.1000115. Acesso em 16 jan 2024.

BIOINFO – Revista Brasileira de Bioinformática e Biologia Computacional. [S.l.], 2022. Disponível em: https://bioinfo.com.br/. Acesso em: 19 ago. 2023.

BOLYEN, E. *et al*. Reproducible, interactive, scalable, and extensible microbiome data science using QIIME 2. Nature Biotechnology, v. 37, n. 8, p. 852–857, 2019. Disponível em: https://doi.org/10.1038/s41587-019-0209-9. Acesso em 16 jan 2024.

CHEN, C.; GRENNAN, K.; BADNER, J.; ZHANG, D.; GERSHON, E.; JIN, L.; LIU, C. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One, v. 6, n. 2, p. e17238, 2011. PMID: 21386892; PMCID: PMC3046121. Disponível em: https://doi.org/10.1371/journal.pone.0017238. Acesso em 16 jan 2024.

CLINIC, Mayo. Breast cancer. Overview, 2022. Disponível em: https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470. Acesso em: 06 out. 2023.

COCHRANE, G.R.; GALPERIN, M.Y. The 2010 Nucleic Acids Research Database Issue and Online Database Collection: a community of data resources. Nucleic Acids Research, v. 38, p. D1-D4, 2010. Disponível em: https://doi.org/10.1093/nar/gkp1077. Acesso em 16 jan 2024.

CONSORTIUM A. I. M. Evidence for network evolution in an Arabidopsis interactome map. Science, v. 333, p. 601–607, 2011. https://doi.org/10.1126/science.1203877. Acesso em 16 jan 2024.

DUTKOWSKI, J.; KRAMER, M.; SURMA, MA.; BALAKRISHNAN, R.; CHERRY, JM.; KROGAN, NJ.; IDEKER, T. A gene ontology inferred from molecular networks. Nat. Biotechnol., v. 31, p. 38–45, 2013. Disponível em: https://doi.org/10.1038/nbt.2463. Acesso em 16 jan 2024.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res., v. 30, n. 1, p. 207-210, 2002. PMID: 11752295; PMCID: PMC99122. Disponível em: https://doi.org/10.1093/nar/30.1.207. Acesso em 16 jan 2024.

GARIJO, D.; KINNINGS, S.; XIE, L.; ZHANG, Y.; BOURNE, PE.; *et al*. Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. PLoS ONE, v. 8, n. 11, p. 80278, 2013. Disponível em: https://doi.org/10.1371/journal.pone.0080278. Acesso em 16 jan 2024.

HOGEWEG, P. Simulating the growth of cellular forms. Simulation, v. 31, p. 90-96, 1978. Disponível em https://doi.org/10.1177/003754977803100305. Acesso em 16 jan 2024.

HOGEWEG, P. The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol, v. 7, n. 3, e1002021, 2011. Disponível em: https://doi.org/10.1371/journal.pcbi.1002021. Acesso em 16 jan 2024.

HOGEWEG, P.; HESPER, B. Interactive instruction on population interactions. Comput Biol Med, v. 8, p. 319-327, 1978. Disponível em: https://doi.org/10.1016/0010-4825(78)90032-X. Acesso em 16 jan 2024.

HULL, D.; WOLSTENCROFT, K.; STEVENS, R.; GOBLE, C.; POCOCK, M.R.; LI, P.; OINN, T. Taverna: a tool for building and running workflows of services. Nucleic Acids Research, v. 34, p. W279-W282, 2006. Disponível em: https://doi.org/10.1093/nar/gkl320. Acesso em 16 jan 2024.

INSTITUTO NACIONAL DE CÂNCER (INCA). Bioinformática e biologia computacional. [S.l.], 12 ago. 2022. Disponível em: https://www.gov.br/inca/pt-br/assuntos/pesquisa/pesquisa-basica-e-experimental/bioinformatica-e-biologia-computacional. Acesso em: 19 out. 2023.

ITO, T.; TASHIRO, K.; MUTA, S.; OZAWA, R.; CHIBA, T.; NISHIZAWA, M.; YAMAMOTO, K.; KUHARA, S.; SAKAKI, Y. Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl Acad. Sci. USA, v. 97, n. 3, p. 1143–1147, 2000. Disponível em: https://doi.org/10.1073/pnas.97.3.1143. Acesso em 16 jan 2024.

JIANG, Y.; LIANG, Y.; WANG, D.; XU, D.; JOSHI, T. Gene expression: A dynamic programming approach to integrate gene expression data and network information for pathway model generation. Bioinformatics, v. 36, n. 1, p. 169-176, 2020. Disponível https://doi.org/10.1093/bioinformatics/btz467. Acesso em 16 jan 2024.

KADADI, A.; AGRAWAL, R.; NYAMFUL, C.; ATIQ, R. Challenges of data integration and interoperability in big data. In: Big Data (Big Data), IEEE International Conference On.

IEEE, p. 38–40, 2014. Disponível em: https://doi.org/10.1109/BigData.2014.7004486. Acesso em 16 jan 2024.

KAUPPINEN, T.; DE ESPINDOLA, GM. Linked open science-communicating, sharing and evaluating data, methods and results for executable papers. Procedia Comput Sci., v. 4, p. 726–31, 2011. Disponível em: https://doi.org/10.1016/j.procs.2011.04.076. Acesso em 16 jan 2024.

MA'AYAN, A.; ROUILLARD, AD.; CLARK, NR.; WANG, Z.; DUAN, Q.; KOU, Y. Lean big data integration in systems biology and systems pharmacology. Trends Pharmacol Sci., v. 35, n. 9, p. 450–60, 2014. Disponível em: https://doi.org/10.1016/j.tips.2014.07.001. Acesso em 16 jan 2024.

MITRA, K.; CARVUNIS, A. R.; RAMESH, S. K.; IDEKER, T. Integrative approaches for finding modular structure in biological networks. Nat. Rev. Genet., v. 14, p. 719-732, 2013. Disponível em: https://doi.org/10.1038/nrg3552. Acesso em 16 jan 2024.

MÖLLER, S.; KRABBENHÖFT, H. N.; TILLE, A.; PALEINO, D.; WILLIAMS, A.; WOLSTENCROFT, K.; GOBLE, C.; HOLLAND, R.; BELHACHEMI, D.; PLESSY, C. Community-driven computational biology with Debian Linux. BMC Bioinformatics, Boston, MA, USA. 9-10 July 2010. Disponível em: https://doi.org/10.1186/1471-2105-11-S12-S5. Acesso em 16 jan 2024.

MORRIS, J. S.; BALADANDAYUTHAPANI, V. Statistical Contributions to Bioinformatics: Design, Modeling, Structure Learning, and Integration. Stat Modelling, v. 17, n. 4-5, p. 245-289, 2017. Disponível em: https://doi.org/10.1177/1471082X17698255. Acesso em: 06 ago 2023.

MOSTAFAVI, S.; RAY, D.; WARDE-FARLEY, D.; GROUIOS, C.; MORRIS, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol., v. 9, Suppl 1, p. S4, 2008. Disponível em: https://doi.org/10.1186/gb-2008-9-s1-s4. Acesso em 16 jan 2024.

NATIONAL CANCER INSTITUTE. What is cancer? 2021. Disponível em: https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Acesso em: 06 out. 2023.

NATIONAL INSTITUTES OF HEALTH (NIH). NIH Working Definition of Bioinformatics and Computational Biology. [S.l.], 17 jul. 2000. Disponível em https://web.archive.org/web/20190430234827id_/http://www.binf.gmu.edu/jafri/math6390-bioinformatics/workingdef.pdf. Acesso em: 06 ago 2023

NEKRUTENKO, A.; TAYLOR, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet., v. 13, n. 9, p. 667–72, 2012. Disponível em: https://doi.org/10.1038/nrg3305. Acesso em 16 jan 2024.

PROSDOCIMI, F. Introdução à bioinformática. Capítulo 1: Uma visão global da bioinformática. [S.l.], 2007. Disponível em: https://professor.pucgoias.edu.br/SiteDocente/admin/arquivosUpload/18497/material/Cap.%201%20Vis%C3%A3o%20Global%20da%20Bioinform%C3%A1tica.pdf. Acesso em: 28 jun. 2023.

PRUITT, KD.; TATUSOVA, T.; BROWN, GR.; MAGLOTT, DR. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res., v. 40, p. D130–D135, 2012. Disponível em: https://doi.org/10.1093/nar/gkr1079. Acesso em 16 jan 2024.

RSTUDIO CLOUD. RStudio Cloud: Do, Share, Teach, and Learn Data Science. Disponível em: https://rstudio.cloud/. Acesso em: 21 jul. 2023.

SALEEM, M.; KAMDAR, MR.; IQBAL, A.; SAMPATH, S.; DEUS, HF.; NGOMO, A-CN. Big linked cancer data: Integrating linked TCGA and PubMed. Web Semant Sci Serv Agents World Wide Web, v. 27, p. 34–41, 2014. Disponível em: https://doi.org/10.1016/j.websem.2014.07.004. Acesso em 16 jan 2024.

SOLIMAN, J.; BRAZIER, Y.; VILLINES, Z. Colorectal cancer: Symptoms, treatment, risk factors, and causes, 2023. Disponível em: https://www.medicalnewstoday.com/articles/155598. Acesso em: 06 out. 2023.

THE ECONOMIST. Drowning in data. [S.l.], 26 jun. 1999. Disponível em: https://www.economist.com/science-and-technology/1999/06/24/drowning-in-data. Acesso em: 06 ago 2023.

TULLY, B.J.; BUONGIORNO, J.; COHEN, A.B.; CRAM, J.A.; GARBER, A.I.; HU, S.K.; KRINOS, A.I.; LEFTWICH, P.T.; MARSHALL, A.J.; SIERADZKI, E.T.; SPETH, D.R.; SUTER, E.A.; TRIVEDI, C.B.; VALENTIN-ALVARADO, L.E.; WEISSMAN, J.L. The Bioinformatics Virtual Coordination Network: An Open-Source and Interactive Learning Environment. Frontiers in Education, 2021. v. 6, p. 1-10. Disponível em: https://doi.org/10.3389/feduc.2021.711618. Acesso em 16 jan 2024.

WANDELT, S.; RHEINLÄNDER, A.; BUX, M.; THALHEIM, L.; HALDEMANN, B.; LESER, U. Data management challenges in next-generation sequencing. Datenbank-Spektrum, v. 12, n. 3, p. 161–71, 2012. Disponível em: https://doi.org/10.1007/s13222-012-0098-2. Acesso em 16 jan 2024.

WELCH, L. R.; SCHWARTZ, R.; LEWITTER, F. A report of the Curriculum Task Force of the ISCB Education Committee. PLOS Comput Biol, v. 8, e1002570, 2012. Disponível em: https://doi.org/10.1371/journal.pcbi.1002570. Acesso em 16 jan 2024.

WELCH, L.; LEWITTER F.; SCHWARTZ R.; BROOKSBANK C.; RADIVOJAC P.; GAETA B., *et al*. Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLoS Comput Biol, v. 10, n. 3, e1003496, 2014. Disponível em: https://doi.org/10.1371/journal.pcbi.1003496. Acesso em 16 jan 2024.
.

WORLD HEALTH ORGANIZATION. Cancer, 2023. Disponível em: https://www.who.int/health-topics/cancer#tab=tab_1. Acesso em: 06 out. 2023.

WORLD HEALTH ORGANIZATION. Colorectal cancer, 2023. Disponível em: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer. Acesso em: 06 out. 2023.

YAO, R.; QIAN, J.; HUANG, Q. Deep-learning with synthetic data enables automated picking of cryo-EM particle images of biological macromolecules. Bioinformatics, v. 36, n. 4, p. 1252-1259, Feb. 2020. Disponível em: https://doi.org/10.1093/bioinformatics/btz728. Acesso em 16 jan 2024.