Moléculas da Amazônia: banco de dados integrado para apoiar pesquisas em saúde e biotecnologia na Amazônia

Carolina Barros da Costa
Instituto Federal de Rondônia
Campus Calama
RO, Brasil
carolinabc962@gmail.com

Izabela de Santos Albuquerque
Instituto Federal de Rondônia
Campus Calama
RO, Brasil
euizabellaa@gmail.com

Rayssa Cristhiny Santos Baker
Instituto Federal de Rondônia
Campus Calama
RO, Brasil
rayssa.seringueiras@gmail.com

Jean Carlo Wai Keung Ma
Instituto Federal de Mato Grosso do Sul
MS, Brasil
realjcwkm@gmail.com

Márcio Rodrigues Miranda
Instituto Federal de Rondônia
Campus Calama
RO, Brasil
marcio.miranda@ifro.edu.br

Ryan da Silva Ramos
Universidade da Amazônia - UNAMA
AP, Brasil
ryanquimico@gmail.com

Kaio Alexandre da Silva
Instituto Federal de Rondônia
Campus Calama
RO, Brasil
kaio.silva@ifro.edu.br

Abstract

Amazonian biodiversity is widely recognized as one of the richest and most diverse in the world. Covering approximately 40% of Brazilian territory, the Amazonian biome harbors a flora rich in natural resources that can be sustainably exploited to generate new products and innovative processes. Biological databases have been built to aid in a better understanding of these areas, as well as in identifying possible applications of their biological resources. However, the fragmentation and dispersion of data on species, biomolecules, and digital structures of the Amazonian flora across multiple repositories hinders bioprospecting processes, since researchers need to consult various databases to obtain information. Therefore, the objective of this article is to present Moléculas da Amazônia, a database developed to unify information on the species present in the Amazonian flora, as well as data on biomolecules that can be found in these species. For this, Web Scraping and ETL (Extract, Transform and Load) techniques were used to collect, integrate, and store data provided by GBIF, NuBBEDB, NCBI, and ChEMBL. The methodology resulted in the initial mapping of 118 species, 40 of which are endemic, and the cataloging of more than 500 molecules. With this, the "Molecules of the Amazon" project becomes a strategic tool to accelerate the discovery of new drugs and bioproducts. By organizing the chemical knowledge of the forest and its geographic distribution, the initiative supports the bioeconomy and sustainable development, connecting the natural wealth of the Amazon to technological innovation.

Keywords

Banco de Dados, Web Scraping, ETL

1 Introdução

A biologia experimental vem se transformando cada dia mais, graças aos avanços proporcionados pela bioinformática. Dentre esses avanços, tem-se a disponibilização de ferramentas capazes de simular as interações físicas no meio digital, tal como as ferramentas de ancoragem molecular (*molecular docking*) e a dinâmica molecular.

A ancoragem molecular, ou molecular docking, permite simular o encaixe de um ligante em uma estrutura alvo. Isso permite prever a interação entres eles e identificar potenciais moléculas de interesse e suas aplicações [4].

Já a dinâmica molecular proporciona uma análise mais detalhada a respeito dessas interações a partir da simulação do comportamento de moléculas ao longo do tempo. Ademais, esta também leva em consideração fatores externos, como temperatura e pressão [9].

A união dessas técnicas reduz a necessidade de testes físicos extensivos, diminuindo o uso de materiais e economizando tempo e custos associados aos experimentos tradicionais.

Para apoiar essas análises, bancos de dados biológicos têm desempenhado um papel importante na bioinformática. Com isso, pesquisadores da Universidade Estadual Paulista (UNESP) desenvolveram o NuBBEDB (Banco de Dados do Núcleos de Bioensaios, Ecofisiologia e Biossíntese de Produtos Naturais). Este nasceu com o objetivo de fornecer dados químicos e biológicos da biodiversidade brasileira, possibilitando o acesso a diversos compostos naturais presentes na biodiversidade nacional [10]. Assim como a iniciativa do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro que criou o Reflora, um banco de dados biológico que contribui para o melhor entendimento da biodiversidade presente na flora de todo o Brasil, disponibilizando informações essenciais sobre as suas espécies [12].

É nesse cenário de pesquisa e desenvolvimento tecnológico que a Amazônia se destaca como fonte de uma gama de recursos naturais. A região é lar de uma rica biodiversidade, o qual abrange uma população numerosa de espécies de animais, plantas, insetos e microrganismos [11]. As espécies amazônicas têm se tornado cada vez mais um alvo de interesse econômico. Isso ocorre, pois, essa

diversidade biológica e genética abriga matérias-primas básicas para avanços na biotecnologia [1].

No entanto, apesar de iniciativas como o Reflora e o NuBBEDB, não há um banco de dados que centralize e integre informações da flora amazônica que abrangem as espécies, biomoléculas e estruturas digitais, independente dos formatos de arquivos, nível estrutural ou formato de visualização, estando estas dispersas em diferentes bancos de dados. Essa fragmentação limita o potencial de prospecção de compostos naturais, uma vez que o pesquisador necessita consultar múltiplos bancos de dados para obter informações.

Logo, o objetivo deste estudo é apresentar o Moléculas da Amazônia, um banco de dados que centraliza e integra dados da flora amazônica, abrangendo espécies, biomoléculas e estruturas digitais, que atualmente estão dispersos em diferentes bancos de dados.

A integração desses dados tem capacidade de acelerar processos de bioprospecção, permitir uma melhor compreensão do potencial farmacêutico e/ou biotecnológicos de compostos da flora amazônica, promover o desenvolvimento sustentável de novos produtos e/ou processos, e fornecer uma visão abrangente e acessível da biodiversidade amazônica, apoiando a pesquisa, desenvolvimento e inovação (PD&I).

Este artigo está organizado como segue: a Seção 2 a fragmentação de dados entre bancos de dados existentes; a Seção 3 apresenta a solução proposta; a Seção 4 apresenta os resultados; e a Seção 5 apresenta as considerações finais.

2 Fragmentação de dados entre Bancos de Dados Existentes

O acesso a informações de alta qualidade permite que pesquisadores possam obter um conhecimento mais aprofundado sobre a complexidade de sistemas ecológicos, uma vez que tornam possível organizar e analisar grandes volumes de informações obtidas de diferentes fontes e identificar padrões complexos [6].

O GBIF (Global Biodiversity Information Facility) é um banco de dados que surgiu devido a necessidade de tornar dados e informações sobre a biodiversidade acessíveis para o mundo. Estabelecido oficialmente em 2001, trata-se de uma organização internacional responsável por fornecer acesso gratuito a dados das espécies presentes na biodiversidade global, bem como a ocorrência destas, para a promoção da pesquisa científica [3].

Outro importante bancos de dados desenvolvido para disponibilizar informações sobre a biodiversidade brasileira, porém como foco em espécies da sua flora e funga, é o Reflora. Criado em 2010, o programa Reflora nasceu com o objetivo de resgatar, por meio de imagens de alta resolução, espécies da flora brasileira presentes em herbários estrangeiros. A partir disso, foram criados dois programas: o Herbário Virtual Reflora, para receber, armazenar e publicar imagens, e o Flora e Funga do Brasil, para validar os nomes atribuídos às imagens presentes no Herbário Virtual Reflora [12].

O NuBBEDB também é um importante projeto desenvolvido para auxiliar na compreensão da biodiversidade brasileira. Porém, ao invés de fornecer informações sobre as espécies, este objetiva disponibilizar dados químicos e biológicos presentes na biodiversidade do Brasil, possibilitando acessar dados de diversos compostos naturais que podem ser encontrados em território nacional [10].

O Lottus é um banco de dados que gerencia conhecimento aberto em pesquisa de produtos naturais e permite que usuários realizem buscas por estruturas moleculares e/ou orientada à taxonomia. Este abriga um projeto de código aberto responsável por armazenar, pesquisa e analisar produtos naturais (NPs) [7].

O ChEMBL é um banco de dados que disponibiliza informações sobre a bioatividade e propriedades químicas de moléculas [2]. Este permite que pesquisadores obtenham informações sobre as características de uma molécula de interesse e, para aprofundar seu conhecimento, este podem associar essas informações com os dados biomédicos e genômicos disponibilizados pelo NCBI (National Center for Biotechnology Information) [5].

Mediante esta variedade de repositórios, o Moléculas da Amazônia está sendo desenvolvido de modo a oferecer uma solução unificada, que centraliza e integra dados taxonômicos, distribuição geográfica, dados moleculares, dados bibliográficos e dados patentários sobre os recursos naturais provenientes da flora amazônica.

3 Solução Proposta

Este trabalho propõe o desenvolvimento do Moléculas da Amazônia, o qual seguiu uma adaptação metodológica no Cross-Industry Standard Process for Data Mining (CRISP-DM) com um pipeline estruturado com base no modelo ETL (Extract, Transform, Load) reprodutível [8].

O pipeline foi organizado de forma modular, permitindo a reexecução integral do processo, auditoria das etapas e rastreabilidade das transformações aplicadas aos dados. Todas as etapas foram automatizadas por meio de scripts versionados, garantindo consistência e reprodutibilidade dos resultados.

A etapa de extração consistiu na coleta automatizada de dados a partir de portais públicos, bases institucionais e repositórios digitais, utilizando técnicas de Web Scraping e, quando disponíveis, APIs REST.

Para alimentar o Moléculas da Amazônia, foram utilizados os dados contidos nos repositórios: GBIF, NuBBEDB, NCBI e ChEMBL. O repositório do NuBBEDB forneceu dados químicos e biológicos de produtos naturais da biodiversidade brasileira já identificados e catalogados de acordo com a espécie da planta.

O repositório do GBIF permitiu o acesso a dados taxonômicos detalhados e distribuição geográfica sobre as espécies da flora amazônica, o que possibilitou a identificação de quais plantas estavam localizadas nos estados da Amazônia Legal Brasileira.

As coletas nos repositórios do NCBI e ChEMBL permitiram acessar dados genéticos e moleculares, patentes solicitadas, proteínas alvo, cuidados e perigos associadas as moléculas e referências bibliográficas associadas.

Para a diagramação do banco de dados, foi utilizada a ferramenta MySQL Workbench. Esta é uma ferramenta visual unificada que permite projetar, modelar, gerar e gerenciar banco de dados visualmente.

A coleta de dados foi realizada com a Linguagem Python e os frameworks Scrapy e Selenium, possibilitando a interação com navegadores, raspagem de dados e a navegação entre repositórios de acordo com o pipeline a ser realizado para a captura e classificação do dado.

4 Resultados

As técnicas de Web Scraping e ETL foram implementadas por meio do framework Scrapy em Python. Onde foram desenvolvidos algoritmos específicos para a execução da coleta de dados diretamente dos sites.

Foi realizada a identificação das páginas e a extração automatizada dos dados utilizando robôs em Python que percorreram as páginas e coletaram os dados de interesse.

A extração de dados automatizada foi iniciada pelo NuBBEDB. A partir de uma biblioteca Scrapy em Python, foi criado um algoritmo que percorreu as páginas do NuBBEDB e coletou as informações referentes às biomoléculas que estão localizadas na flora brasileira e em quais espécies podem ser encontradas.

Com os dados coletados no NuBBEDB, foi feita uma consulta no API do GBIF, com o objetivo de obter informações taxonômicas detalhadas (família, gênero e espécie) e dados geográficos que indicam a ocorrências dessas espécies. Desta forma, foi realizado o filtro selecionando apenas espécies que tiveram a ocorrência registrada em algum estado da Amazônia Legal Brasileira. Identificando quais estados elas foram registradas, e categorizando se são endêmicas e nativas daquela região.

Em seguida, para complementar dados sobre as moléculas, foram feitas consultas no NCBI e ChEMBL. Através do ChEMBL foi possível coletar informações adicionais sobre as propriedades físico-químicas das moléculas, bem como sua bioatividade. Enquanto o NCBI, contribuiu para complementar essas informações com dados genéticos, moleculares, quais patentes foram solicitadas, proteínas alvo, cuidados e perigos associadas as moléculas e referências bibliográficas associadas. Tais consultas permitiram a integração e o armazenamento de dados biológicos e químicos complementares no Moléculas da Amazônia, contribuindo para enriquecer o conjunto de informações disponíveis.

A coleta inicial realizada no NuBBEDB resultou na extração de 572 moléculas. A partir desse resultado, foram coletados dados taxonômicos e geográficos dentro do GBIF. A coleta de dados taxonômicos e geográficos retornou um total de 30 famílias, 72 gêneros e 118 espécies.

Quanto a distribuição geográfica das espécies coletadas, os estados do Mato Grosso, Pará e Amazonas lideram com 57, 56 e 53 espécies, respectivamente. Por outro lado, os estados de Roraima e Amapá registraram o menor número de espécies, com 22 e 30 espécies, respectivamente.

Em relação as moléculas, os resultados condizem com a maior diversidade de espécies da região, com os estados do Mato Grosso, Pará e Amazonas também liderando esse cenário com 314, 294 e 261 moléculas, respectivamente.

A origem dessas espécies foi categorizada em nativas, que ocorrem naturalmente em uma determinada região, e cultivadas, espécies semeadas/plantadas pelo ser humano. Das 118 espécies coletadas, 114 são nativas e 4 são cultivadas.

Além disso, foi possível identificar que 40 espécies são endêmicas, ou seja, são espécies exclusivas da região Amazônica e não ocorrem em nenhum outro lugar do mundo.

Entre as espécies coletadas, aquelas que mais possuem moléculas associadas no banco de dados são: Flacourtiaceae Casearia sylvestris, com 27 moléculas; Rubiaceae Chimarrhis turbinata, com 19 moléculas; Myristicaceae Virola sebifera, com 17 moléculas; Meliaceae Cedrela odorata, com 15 moléculas; e Myristicaceae Iryanthera juruensis, com 15 moléculas.

A disponibilização dessas informações não apenas cataloga a flora amazônica, mas agregam valor tangível à sua biodiversidade. A estruturação detalhada dos perfis químicos das espécies vegetais desta região representa um ativo estratégico fundamental para o ecossistema de pesquisa e inovação na Amazônia.

5 Considerações Finais

Os resultados obtidos com o Moléculas da Amazônia validam a eficácia das técnicas de Web Scraping e ETL aplicadas para a centralização e integração de dados de diferentes fontes, como GBIF, NuBBEDB, NCBI e ChEMBL, em um único local.

A estruturação inicial do banco de dados não apenas permitiu a unificação de informações taxonômicas e moleculares, mas também revelou um potencial significativo para a bioprospecção, evidenciado pelo mapeamento de 118 espécies, das quais 40 são endêmicas, e a catalogação de mais de 500 moléculas.

Com isso, o banco de dados Moléculas da Amazônia não atua somente como um armazém de dados, mas como uma ferramenta de apoio a pesquisa que conecta a riqueza taxonômica da Amazônia às suas propriedades moleculares e farmacológicas, promovendo a bioeconomia e o desenvolvimento sustentável de novos produtos.

As etapas posteriores incluem o desenvolvimento da plataforma digital para a disponibilização do acesso ao banco de dados, a coleta e integração com o repositório de saberes tradicionais tutelado pela Fundação Oswaldo Cruz (Fiocruz) e a coleta e integração com o repositório Lottus.

References

[1] Sarita Albagli. 1998. Da biodiversidade à biotecnologia: a nova fronteira da informação. *Ci. Inf., Brasília* 27, 1 (1998), 7–10.

[2] ChEMBL. 2024. *About.* https://chembl.gitbook.io/chembl-interface-documentation/about

[3] Sistema Global de Informação sobre Biodiversidade GBIF. 2024. *O que é o GBIF?* https://www.gbif.org/pt/what-is-gbif

[4] Jiyu Fan, Ailing Fu, and Le Zhang. 2019. Progress in molecular docking. *Progress in molecular docking* 7, 2 (2019), 83–89. doi:10.1007/s40484-019-0172-y

[5] National Center for Biotechnology Information. 2024. *About NCBI.* https://www.ncbi.nlm.nih.gov/home/about/

[6] Steven Kelling, wesley M. Hochachka, Daniel Fink, Mirek Riedewald, Rich Caruana, Grant Ballard, and Giles Hooker. 2009. Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience* 59, 7 (2009), 613–620. doi:10.1525/bio.2009.59.7.12

[7] Lotus. 2024. *Lotus.* https://lotus.naturalproducts.net/

[8] Gonzalo Mariscal, Óscar Marbán, and Covadonga Fernández. 2010. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review* 25, 2 (2010), 137–166. doi:10.1017/S0269888910000032

[9] A. C. Namba, V. B. da Silva, and C. H. T. P. da Silva. 2008. Dinâmica molecular: teoria e aplicações em planejamento de fármacos. *Eclética Química* 33, 4 (2008), 13–24. doi:10.1590/S0100-46702008000400002

[10] Alan C Pilon, Marilia Valli, Alessandra C Dametto, Meri Emili F Pinto, Rafael T Freire, Ian Castro-Gamboa, Adriano D Andricopulo, and Vanderlan S Bolzani. 2017. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Scientific Reports* 7, 1, Article 7215 (2017). doi:10.1038/s41598-017-07451-x

[11] President Mark J. Plotkin. 2020. *The Amazon: what everyone needs to know.* Oxford University Press, New York, NY.

[12] Reflora. 2025. *Programa Reflora.* https://reflora.jbrj.gov.br/reflora/PrincipalUC/PrincipalUC.do