# BAMBU SYSTEMATIC REVIEW: UMA FERRAMENTA DE EXTRAÇÃO DE INFORMAÇÃO PARA AUXÍLIO À REVISÃO SISTEMÁTICA

GABRIEL LISTON DE MENEK [1]; GRATCHELA DUTRA RODRIGUES [2]; DARLING DE ANDRADE LOURENÇO [3]; FREDERICO SCHMITT KREMER [4]

*[1] Omixlab, Universidade Federal de Pelotas – gabriellistondemenek@gmail.com*
*[2] Omixlab, Universidade Federal de Pelotas – gratirodrigues.gdr@gmail.com,*
*[3] BIOSCIENT; Omixlab, Universidade Federal de Pelotas – darlinglourenco@gmail.com*
*[4] Omixlab, Universidade Federal de Pelotas – fred.s.kremer@gmail.com*

## 1. INTRODUÇÃO

A prospecção de novos fármacos nos permite a identificação e caracterização de moléculas com potencial de modular enfermidades. No entanto, esse ainda é um processo longo e oneroso, podendo levar entre 10-15 anos (Koutroumpa, 2023) para uma molécula suficientemente segura e eficiente chegar ao mercado. Uma de suas principais e mais desafiadoras etapas é a definição dos alvos moleculares em que as moléculas candidatas a fármacos vão ser testadas (Rasul et al., 2022), o que implica a relevância de métodos para a identificação dos alvos moleculares que estejam relacionados a alguma enfermidade. Dentre as opções disponíveis atualmente, a busca sistemática por alvos moleculares na literatura científica é uma forma interessante de identificar alvos (Singh, 2020).

A revisão sistemática de artigos envolve uma busca utilizando métodos reprodutíveis para encontrar, selecionar e sintetizar todas as evidências disponíveis, de acordo com critérios pré-estabelecidos (Gopalakrishnan, 2013). Ela é uma forma de facilitar o acesso às evidências de múltiplos estudos de forma eficiente, permitindo uma análise dos diversos resultados que não seria possível por meio da leitura dos artigos individualmente (Page et al., 2021), sendo muito mais rápida, e também, mais segura, visto que visa reduzir os vieses inseridos na pesquisa. No entanto, seu desenvolvimento também é um processo longo, podendo levar, de acordo com (O'dwyer, 2021), entre 6 a 18 meses para ser finalizado. Além disso, normalmente, é necessário uma equipe para sua realização, visto que, envolve o estabelecimento de critérios de inclusão e exclusão, a busca nas bases de dados, a triagem dos artigos, extração de seus dados e anotação dos resultados (Uman, 2011).

Neste trabalho, apresentamos o desenvolvimento do *Bambu Systematic Review*, uma plataforma para a triagem de artigos, extraindo informações como genes e alvos moleculares relevantes de uma forma automatizada, permitindo aos autores um método mais eficiente para o desenvolvimento de revisões sistemáticas, economizando o tempo de realização das etapas iniciais e entregando as informações necessárias de uma forma já sumarizada.

## 2. METODOLOGIA

O *Bambu Systematic Review* foi proposto no contexto das demandas da startup BIOSCIENT, que atualmente é conveniada à UFPEL e ao Omixlab. A aplicação foi desenvolvida utilizando a linguagem de programação Python, na versão 3.8, e o pacote Conda (https://conda.io/) foi utilizado para o gerenciamento

de ambientes virtuais e bibliotecas. Essa ferramenta trabalha com artigos dos principais bancos de dados de literatura científica, sendo eles PubMed, Scopus e ScienceDirect. Na etapa inicial, a extração de artigos destas bases de dados, foram utilizadas duas bibliotecas: Metapub (https://github.com/metapub/metapub), para o PubMed, e Elsapy (https://github.com/ElsevierDev/elsapy) para o Scopus e ScienceDirect, que pertencem à Elsevier. Para cada uma das opções de bancos de dados disponíveis na plataforma, foi criada uma função que recebe os parâmetros de busca, que são as palavras-chaves, formatadas com os filtros e operadores booleanos, e o número de artigos. A função executa uma busca no seu respectivo banco de dados com as palavras-chaves escolhidas, extraindo informações dos artigos, como título, resumo, autores, páginas, revista onde foi publicado, tipo de artigo, *Digital Object Identifier* (DOI), afiliações e outras informações específicas do banco de dados.

Após isso, a função retorna os resultados no formato de um *dataframe*, construído em um objeto do Pandas (https://github.com/pandas-dev/pandas), uma biblioteca de manipulação e análise de dados em python. Assim que a busca for concluída em todos os repositórios, os três *dataframes* são formatados com uma indexação padrão e concatenados em um único *dataframe*, dessa forma, as informações extraídas dos três bancos de dados diferentes podem ser visualizadas em um único arquivo de forma organizada e unificada. No entanto, o conteúdo do *dataframe* de resultados por si só não mostra muito mais do que uma busca padrão nos repositórios. Para enriquecê-lo, algumas bibliotecas podem ser úteis com o objetivo de complementar os resultados com informações relevantes ao pesquisador na produção de uma revisão sistemática.

A primeira biblioteca utilizada para esse fim foi o Spacy (https://github.com/explosion/spaCy), que utilizando o modelo BIONLP13CG do SciSpacy (https://github.com/allenai/scispacy), é capaz de identificar, termos relacionados à biologia, dentre eles, os genes e produtos de genes. Este modelo foi utilizado para extrair os genes citados em cada artigo, no entanto, os produtos dos genes geram uma informação indesejada na visualização. Sendo assim, a biblioteca FlashText (https://github.com/vi3k6i5/flashtext) foi utilizada para treinar um modelo com os nomes de genes humanos registrados no National Center for Biotechnology Information (NCBI), o qual foi utilizado para filtrar apenas os genes extraídos pelo SciSpacy. Por fim, os genes citados em cada artigo foram adicionados em uma nova coluna no *dataframe* dos resultados.

Para o desenvolvimento front-end da plataforma, foi utilizado o Flask https://github.com/pallets/flask), esse é um *framework* simples e flexível, com diversas bibliotecas de extensões disponíveis. A aplicação básica de Flask funciona por um sistema de rotas, onde cada rota acessada executa o código de uma função em python, retornando uma página em HTML para o usuário. Esse sistema não traz consigo os elementos necessários para a interface de pesquisa da plataforma, para isso, foram utilizadas as bibliotecas WTForms (https://github.com/wtforms/wtforms) e Flask-WTF (https://github.com/wtforms/flask-wtf) para a renderização e validação de formulários. Estes formulários foram utilizados tanto para as ferramentas de login e registro de usuário, quanto de pesquisa por artigos na plataforma.

Para a autenticação de usuários, foi utilizada a biblioteca Flask-Login (https://github.com/maxcountryman/flask-login), que permite um gerenciamento de

sessão do usuário, limitando o acesso apenas às páginas que a conta logada possui permissão. Quando uma nova conta é criada, suas credenciais são enviadas para o banco de dados da plataforma, onde ficarão armazenadas junto dos resultados de suas pesquisas. O banco de dados é gerenciado com uso da biblioteca Flask-SQLAlchemy (https://github.com/pallets-eco/flask-sqlalchemy/) e nele a senha é criptografada usando a biblioteca Bcrypt (https://github.com/pyca/bcrypt/). Além disso, o processamento de cada pesquisa é executado de forma assíncrona utilizando o Celery (https://github.com/celery/celery), uma biblioteca capaz de organizar uma fila de tarefas, desta forma o usuário pode continuar navegando, e até realizar mais pesquisas, enquanto sua solicitação é processada.

## 3. RESULTADOS E DISCUSSÃO

Atualmente, o Bambu Systematic Review possui toda a estrutura básica para a extração de artigos, além da autenticação de usuários. Em todas as páginas, está presente uma aba de navegação, nela estão disponíveis as opções de voltar à página inicial ou acessar a pesquisa de artigos. Caso o usuário não esteja logado, receberá também as opções de fazer login ou registrar uma nova conta; caso esteja logado, poderá acessar sua área do usuário ou realizar logout. Esta aba é o único elemento funcional presente na página inicial, que serve apenas como um ponto de partida. Na página de login, o usuário pode digitar suas credenciais ou, caso ainda não possua uma conta, acessar a página de registro para criar uma. Quando uma nova conta é criada, seus dados são armazenados no banco de dados da aplicação, com a senha sendo criptografada.

Após o login, a pesquisa de artigos possui as interfaces básica e avançada. A primeira é mais simples, mas limitada apenas às opções de pesquisa que ambos os repositórios possuem em comum. A avançada possui opções de pesquisa específicas para cada repositório, mas exige que o usuário adicione os termos de busca separadamente para cada repositório. Nelas são inseridos o termo de busca, o campo de busca, e o operador booleano. Com essas informações são construídas duas “queries” diferentes, uma para o PubMed e outra para o Elsevier. Após isso, o usuário pode escolher de quais repositórios deseja extrair os artigos e quantos artigos extrair, com um limite de 5000 artigos por pesquisa.

Quando finalizada, o resultado da pesquisa será salvo em formato de *comma-separated values* (CSV), e poderá ser acessado por uma lista de resultados na página da área do usuário, com as opções de visualizar o resultado no próprio navegador ou baixar seu arquivo. Os resultados fornecem uma tabela que organiza as informações extraídas de cada artigo, sendo elas: Título, resumo, páginas, revista da publicação, autores, data, tipo de publicação, DOI, afiliações, MeSH Terms e, opcionalmente, genes encontrados no resumo.

Previamente, outras ferramentas semelhantes já foram publicadas, no entanto, cada uma delas possui suas limitações. O litstudy (https://github.com/ElsevierSoftwareX/SOFTX-D-22-00050) e litreviewer (https://github.com/Kamakshaiah/literature-review) são duas ferramentas de revisão de artigos limitadas a pesquisas exclusivamente no Scopus e Google Scholar, respectivamente. O ASReview (https://github.com/asreview/asreview) é

uma ferramenta que classifica os artigos por sua relevância usando *machine learning*, mas a busca de artigos não é automatizada. O bibliometrix (https://www.bibliometrix.org/home/) é uma ferramenta desenvolvida em R que, para realizar suas análises, é necessário que o usuário possua os artigos salvos em uma lista BibTeX.

## 4. CONCLUSÕES

Essa foi a primeira versão do *Bambu Systematic Review*, que já possui todas as suas funções essenciais para a extração de artigos e organização dos resultados em *dataframes*. Sendo assim, a aplicação fornece, de forma automatizada, uma sumarização das informações de artigos científicos que pode ser útil para o pesquisador na escrita de análises sistemáticas de uma forma mais eficiente. De acordo com o caráter inovador da aplicação desenvolvida, o pedido de registro de software já foi submetido. Futuramente, visamos aprimorar as funcionalidades já disponíveis e implementar novas bibliotecas e fontes de dados que possam ser úteis na extração de informações dos artigos, incluindo o uso de *large language models* (LLMs), como o ChatGPT e LLaMA, para extração de dados a partir de abstracts e textos inteiros dos manuscritos de forma estruturada. Por fim, Para aprimorar as funcionalidades atuais, planejamos adicionar mais opções para o reconhecimento de termos científicos baseado no SciSpacy, treinar novos modelos do FlashText para serem usados na pesquisa de outros genes além dos humanos, e, até mesmo, uma funcionalidade de treinamento de modelos personalizados para que o usuário possa buscar por genes mais específicos.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS

GUIDOTTI, I. L. et al. Bambu and its applications in the discovery of active molecules against melanoma. **Journal of Molecular Graphics and Modelling**, v. 124, p. 108564, 1 nov. 2023.

GOPALAKRISHNAN, S.; GANESHKUMAR, P. Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. **Journal of Family Medicine and Primary Care**, v. 2, n. 1, p. 9–14, jan. 2013.

KOUTROUMPA, N.-M. et al. A Systematic Review of Deep Learning Methodologies Used in the Drug Discovery Process with Emphasis on In Vivo Validation. **International Journal of Molecular Sciences**, v. 24, n. 7, p. 6573, jan. 2023.

O'DWYER, L. C.; WAFFORD, Q. E. Addressing challenges with systematic review teams through effective communication: a case report. **Journal of the Medical Library Association : JMLA**, v. 109, n. 4, p. 643–647, 2021.

PAGE, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **Systematic Reviews**, v. 10, n. 1, p. 89, 29 mar. 2021.

RASUL, A. et al. Target Identification Approaches in Drug Discovery. Em: SCOTTI, M. T.; BELLERA, C. L. (Eds.). **Drug Target Selection and Validation**. Computer-Aided Drug Discovery and Design. Cham: Springer International Publishing, 2022. p. 41–59.

SINGH, D. B. (ED.). **Computer-Aided Drug Design**. Singapore: Springer, 2020.

UMAN, L. S. Systematic Reviews and Meta-Analyses. **Journal of the Canadian Academy of Child and Adolescent Psychiatry**, v. 20, n. 1, p. 57–59, fev. 2011.