UNIVERSIDADE DE RIBEIRÃO PRETO - UNAERP
DEPARTAMENTO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

EVERTON RAFAEL DA SILVA

BLASTER: UMA FERRAMENTA DE BIOINFORMÁTICA

RIBEIRÃO PRETO
2020

EVERTON RAFAEL DA SILVA

BLASTER: UMA FERRAMENTA DE BIOINFORMÁTICA

Dissertação apresentada ao Programa de Pós-Graduação *Stricto Sensu* da Universidade de Ribeirão Preto - UNAERP, como parte das exigências para a obtenção do título de Mestre em Biotecnologia.

Orientadora:
Profa. Dra. Sonia Marli Zingaretti.

RIBEIRÃO PRETO

2020

Ficha catalográfica preparada pelo Centro de Processamento
Técnico da Biblioteca Central da UNAERP
- Universidade de Ribeirão Preto -

S856b Blaster: uma ferramenta de Bioinformática / Everton Rafael da Silva. - - Ribeirão Preto, 2020.
101 f.: il.

Orientadora: Profª. Drª. Sonia Marli Zingaretti.

Dissertação (mestrado) - Universidade de Ribeirão Preto, UNAERP, Biotecnologia. Ribeirão Preto, 2020.

1. DNA. 2. RNA. 3. Aminoácidos. 4. Software. I. Título.

CDD 660.6

EVERTON RAFAEL DA SILVA

BLASTER: UMA FERRAMENTA DE BIOINFORMÁTICA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade de Ribeirão Preto, para obtenção do título de Mestre em Biotecnologia.

Área de Concentração: Biotecnologia Aplicada à Saúde

Data da defesa: 06 de agosto de 2020

Resultado: Aprovado

BANCA EXAMINADORA

Profa. Dra. Sonia Mari Zingaretti
Universidade de Ribeirão Preto – UNAERP

Prof. Dr. Dávid Buzatto
Instituto Federal de São Paulo – IFSP

Prof. Dr. Pablo Rodrigo Sanches
Universidade de Ribeirão Preto – UNAERP

RIBEIRÃO PRETO
2020

Quero dedicar esta dissertação à minha orientadora Dra. Sônia Marli Zingareti cuja dedicação e paciência serviram como pilares de sustentação para a conclusão deste trabalho. Dedico também aos meus parceiros de trabalho e amigos Dr. David Buzatto e Dr. Breno Lisi Romano pelo incentivo e persistência em não me deixar desistir, a meus pais que me ensinaram a nunca desistir e principalmente a minha esposa Erica e minha filha Lívia que suportaram a distância, entenderam as necessidades e superaram juntas minha ausência durante esta fase.

Grato por tudo.

AGRADECIMENTOS

Agradeço a todos que me ajudaram e incentivaram nesta fase de minha vida, em principal o Instituto Federal de São Paulo Campus São João da Boa Vista pelo apoio financeiro sem o qual não seria possivel desenvolver este trabalho.

"*O sucesso é uma consequência e não um objetivo*".

Gustave Flaubert

RESUMO

Em função do considerável aumento no número de sequências de nucleotídeos obtidas, oriundas dos diversos sequenciamentos gênicos seja de DNA ou RNA e das pesquisas buscando identificação da função de cada gene, faz-se necessário a utilização de *softwares* que auxiliam os pesquisadores na busca por sequências já conhecidas. Este estudo tem como objetivo o desenvolvimento de uma ferramenta on-line que agregue os *softwares* utilizados para análise de sequências de DNA, RNA e proteínas, retornando ao pesquisador com o auxílio de apenas uma ferramenta, as sequências já conhecidas e armazenadas nas bases de dados do NCBI. Também tem a finalidade de identificar a função das proteínas analisadas, artigos e textos relacionados. Esta ferramenta integra os serviços dispersos de dois locais diferentes, o Servidor do NCBI e o Servidor do UniProt. Os resultados obtidos com as pesquisas de nucleotídeos e aminoácidos com a ferramenta desenvolvida, são idênticos aos obtidos acessando o site do NCBI e também retorna a função das proteínas e dados referentes aos artigos e textos relacionados a pesquisa, provindas do servidor do UniProt. O *software* desenvolvido conta com um sistema automatizado, o qual é executado com os dados inseridos pelo usuário e os resultados são compilados em um único arquivo no formato PDF, facilitando assim futuras pesquisas.

Palavras-chave: DNA, RNA, aminoácidos, *software*.

ABSTRACT

Due to the significant increase in the number of nucleotide sequences, originating from several genetic sequencers, be it DNA or RNA and the search for function identification of each gene, it is necessary to use auxiliary research software or search for sequences already carried out. This study aims to develop an online tool that aggregates the software used to analyze DNA, RNA and protein sequences, returning to the researcher or the help of just one tool, such as sequences already stored and stored in the databases. NCBI data. It also has the possibility to identify the functions of proteins, articles and related texts. This tool integrates the services dispersed in two different locations, the NCBI Server and the UniProt Server. The results with searches for nucleotides and amino acids with a developed tool, are accessible to access the NCBI website and also return a function of the built proteins related to articles and texts related to a research, proved by the server ofUniProt. The developed software contains an automated system, or the data entered by the user is executed and the results are compiled in a single PDF file format, thus facilitating searches.

Keywords: DNA, RNA, protein, software.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AJAX	*Asynchronous Javascript and XML*
BLAST	*Basic Local Alignment Search Tool*
DNA	*Deoxyribonucleic Acid*
GB	*Gigabyte*
GHZ	*Gigahertz*
HTML	*Hypertext Marckup Language*
IFSP-SBV	Instituto Federal de São Paulo Campus São João da Boa Vista
JQuery	Biblioteca JavaScript
MAFFT	*Multiple Alignment Using Fast Fourier Transform*
MEGA	*Molecular Evolutionary Genetics Analysis*
MUSCLE	*Multiple Sequence Comparison by Log-Expectation*
NBFR	*National Biomedical Research Foundation*
NCBI	*National Center of Biotechnology Information*
PHP	*Hypertext Preprocessor*
PIR	*Protein Information Resource*
RAM	*Read Access Memory*
RNA	*Ribonucleic Acid*
SGBD	Sistema de Gerenciamento de Banco de Dados
SQL	*Structured Query Language*
UML	*Unified Modeling Language*
UNAERP	Universidade de Ribeirão Preto
URL	*Uniform Resource Locator*
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Em 1953, Francis Crick e James Watson apresentaram um modelo de uma estrutura tridimencional de dupla hélice, o *Deoxyribonucleic Acid* (DNA). Esta molécula é a mais estudada no mundo, devido ao seu importante papel para os seres vivos. O sequenciamento do DNA tornou-se uma prática útil em todas as áreas biológicas como: estudos evolutivos e filogenéticos das espécies; clonagem gênica, entre outras .

Na década de 1990 com o início do Projeto Genoma Humano, foi iniciado uma cadeia de sequenciamentos de DNA, com o objetivo de identificar, armazenar e desenvolver ferramentas para a análise dos dados de genes humanos e de outros orcanismos vivos. As ferramentas análise tiveram que ser desenvolvidas para auxiliar no gerenciamento de todo este volume de dados que era depositado com frequência nos bancos de dados (GenBank por exemplo).

Com o crescimento da geração de dados oriundos destes sequenciamentos de DNA e posteriormente *Ribonucleic Acid* (RNA), houve a necessidade de criar ferramentas que conseguissem fazer análise de fragmentos dentro das sequências depositadas. A identificação destes dados e a atribuição de função à essas sequências depende da comparação de sequências já depositadas nos bancos de dados.

Neste momento nascia uma nova tecnologia, a bioinformática. Esta tecnologia que utilizava conhecimentos de informátia e programação de computadores, foi apresentada ao público, sendo considerada como tecnologia às áreas científicas e à evolução da ciência.

Devido a velocidade na obtenção de dados de sequenciamentos, vários programas de computador foram desenvolvidos. Alguns *softwares* utilizados para fazer alinhamentos de sequências são o ClustalW, Clustal Omega, MUSCLE, T-Coffee, MAFFT e BLAST todos eles estão disponíveis publicamente para consulta dos pesquisadores. Estes *softwares* são ferramentas que agilizam o processo de análise, sendo possível obter informações de determinadas sequências, ou mesmo, não obter informação alguma, mostrando assim que não há um alvo, podendo ser sequência ainda não identificada nos bancos de dados.

Cada uma destas aplicações tem funções específicas, retornando dados de forma distinta e assim, este trabalho tem a proposta de desenvolver um *software* que agregue ferramentas em um único local, não havendo a necessidade de o usuário executar tarefas em diferentes *softwares*.

A ferramenta desenvolvida utiliza o BLAST como motor de pesquisa e o NCBI como base de dados, onde são feitas consultas com os dados submetidos pelo pesquisador. Ainda no *software* proposto, há uma funcionalidade que permite exibir os artigos e textos relacionados aos dados retornados na busca. Outra funcionalidade é a execução de forma automatizada, em que o usuário inicia a busca com uma sequência nucleotídica ou de aminoácidos e os processos

são executados de forma automática, ao final exibe ao usuário um arquivo *Portable Document Format* (PDF) consolidado com todos os dados obtidos com a busca, como genes identificados, similaridade com outros organismos e funções preditas.

2 REVISÃO DA LITERATURA

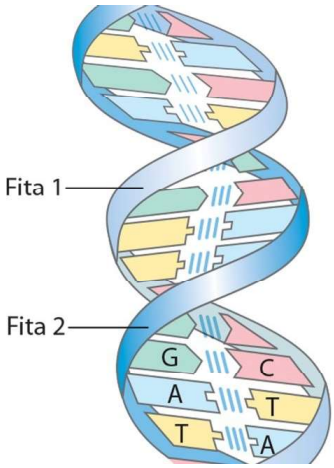
2.1 *ÁCIDO DESOXIRRIBONUCLEICO*

De acordo com Watson *et al.* (2015), o Ácido Desixirribonucleico (*Deoxyribonyucleic Acid* - DNA) é um composto orgânico, cujas moléculas contém as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos. O DNA é composto de pentoses, radicais fosfatos e bases nitrogenadas que formam os nucleotídeos: guanina, citosina, adenina e timina, representadas respectivamente pelas letras G, C, A e T.

Existem cinco tipos de bases nitrogenadas, sendo divididas em dois grupos as purinas e as pirimidinas, sendo que adenina e guanina são purinas e citosina, timina e uracila do grupo pirimidina. A uracila é encontrada somente nas moléculas de Ácido Ribonucleico (*Ribonucleic Acid* - RNA) (ALBERTS *et al.*, 2002).

Na Figura 1, apresenta-se a representação gráfica de uma molécula de DNA, as ligações entre as bases nitrogenadas acontecem por meio de pontes de hidrogênio, sendo que, a adenina (uma purina) se liga com a timina (uma pirimidina) com duas pontes de hidrogênio e a citosina se liga a guanina por três pondes de hidrogênio (WATSON *et al.*, 2015).

Figura 1 – DNA - Dupla Hélice



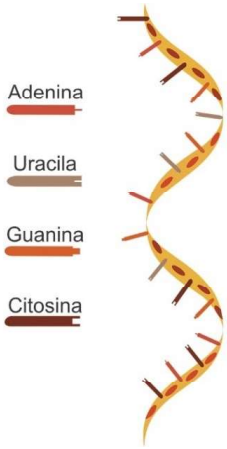
Fonte: Adaptação de (NELSON; COX, 2018), pag. 61.

2.2 ÁCIDO RIBONUCLEICO

Uma das funções do Ácido Ribonucleico (RNA - *Ribonucleic Acid*) é servir de modelo para a síntese das proteínas nas células. Ao contrário do DNA, ele é formado apenas de uma fita simples, que é produzida no núcleo celular a partir de uma das fitas do DNA (transcrição) e suas

bases nitrogenadas são a adenina, a guanina, a citosina e a uracila (WATSON *et al.*, 2015), como pode ser visto na Figura 2.

Figura 2 – RNA - Fita Simples



Fonte: Adaptação de (NELSON; COX, 2018), pag. 1058.

Existem três tipos básicos de RNA: o RNA ribossômico representado pela sigla RNAr (ou rRNA), o RNA mensageiro representado por RNAm (ou mRNA) e o RNA transportador representado por RNAt (ou tRNA) (NELSON; COX, 2018).

- O RNA ribossômico é o de maior peso molecular, fazendo parte da constituição do ribossomo, sendo a organela celular responsável pela síntese de proteínas na célula;
- O RNA mensageiro tem peso molecular intermediário, é responsável por levar a informação do DNA do núcleo até o citoplasma. A partir dessas informações, o RNA mensageiro irá determinar quais são os aminoácidos necessários para a formação de determinada proteína. Este RNA atua em conjunto com o ribossomo na síntese de proteínas;
- E o RNA transportador é o mais leve em peso molecular e sua função é transportar os aminoácidos até os ribossomos que serão utilizados na síntese de proteínas (NELSON; COX, 2018).

Alguns fatores diferenciam o DNA do RNA, tamanho, complexidade, composição e a localização, onde o DNA é encontrado apenas no núcleo das células e o RNA embora seja produzido no núcleo, migra para o citoplasma, onde participa da síntese de proteínas. Outra característica é que no RNA existe a ribose[1] e no DNA desoxirribose[2] (ALBERTS *et al.*, 2002).

Como apresentado nas Seções 2.1, o DNA possui dupla hélice, enquanto o RNA possui uma única fita simples, portanto, o DNA sempre será maior que o RNA.

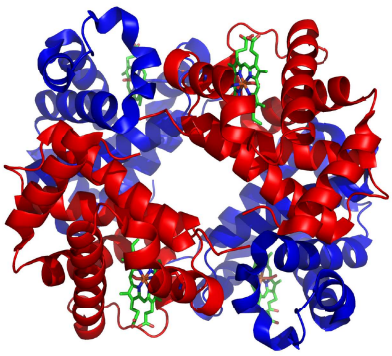
[1] Pentose (açúcar com cinco átomos de carbono) que entra na composição dos ácidos nucleicos.
[2] Pentose de fórmula derivada da ribose por meio da redução da hidroxila no segundo átomo de carbono.

2.3 PROTEÍNAS

As proteínas são constituídas por aminoácidos que formam cadeias complexas por intermédio de ligações peptídicas. Para uma cadeia de aminoácidos ser considerada uma proteína, deve-se observar seu tamanho que não deve ser menor que setenta aminoácidos, se for menor que isso o termo correto a ser utilizado é peptídeo (WATSON *et al.*, 2015).

As proteínas desempenham funções tais como transporte de oxigênio (hemoglobina), anticorpos, catalizadoras de reações químicas (enzimas), receptora em membranas, atuação na contração muscular (actina e miosina), além de serem fundamentais para o crescimento e formação dos hormônios (WATSON *et al.*, 2015). A Figura 3 exemplifica uma proteína Hemoglobina.

Figura 3 – Proteína - Hemoglobina



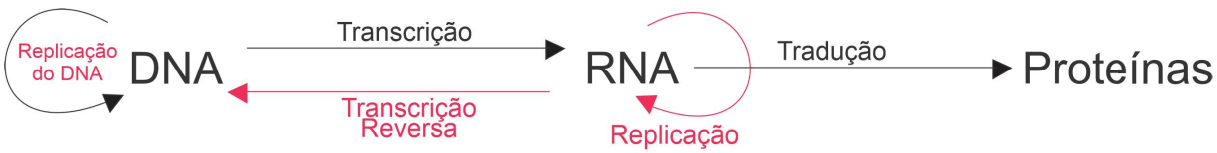
Fonte: Adaptação de (NELSON; COX, 2018), pag. 163.

2.4 DOGMA CENTRAL DA BIOLOGIA MOLECULAR

O Dogma Central da Biologia Molecular é um conceito que ilustra os mecanismos de transmissão e expressão da hereditariedade explicando o fluxo de informação do código genético. O fluxo da informação genética inicia-se no DNA, após ocorre uma transcrição para RNA. Por fim, uma tradução sintetizando a proteína (WATSON *et al.*, 2015).

O dogma foi proposto por Francis Crick em 1958 (CRICK, 1958) e divulgado em um artigo na revista Nature em 1970 (CRICK, 1970), podendo ser visto na Figura 4.

Figura 4 – Dogma Central da Biologia Molecular



Fonte: Adaptação de (NELSON; COX, 2018), pag. 163.

Ainda na Figura 4, a seta que circunda o DNA significa que ele é o molde para a sua própria replicação[3]. A transcrição (representada por uma seta entre o DNA e o RNA) é a primeira etapa da expressão gênica, onde a enzima RNA Polimerase interpreta a sequência de DNA, produzindo uma cadeia de RNA antiparalela complementar. Da mesma forma a síntese de protenínas (chamada de tradução, processo biológico no qual a sequência nucleótica de uma molécula de mRNA é utilizada para ordenar a síntese de uma sequência de aminoácidos determina uma proteína) é coordenada por um molde de RNA. Transcrição reversa é o processo contrário, pois o DNA é sintetizado a partir do RNA (comum em vírus) (WATSON *et al.*, 2015).

2.5 BIOINFORMÁTICA

De acordo com El-Metwally, Ouda e Helmy (2014) a bioinformática é a disciplina que aplica os princípios da ciência da computação, matemática e engenharia para responder questões relacionadas a biologia. O termo bioinformática define uma área da ciência que usa programas de computadores para construir modelos das moléculas que compõem os seres vivos, desde o DNA, RNA e proteínas (LESK, 2019).

A comparação de sequências de nucleotídeos e aminoácidos em bioinformática é largamente utilizada por profissionais nas áreas de biologia molecular, genética e bioquímica. Estas comparações possibilitam a descoberta de novas sequências de DNA que codificam proteínas. A comparação de sequências pode proporcionar descobertas como antepassados de organismos, árvores filogenéticas[4], estruturas e funções de proteínas e estruturas genéticas (DNA) (SETUBAL; MEIDANIS; SETUBAL-MEIDANIS, 1997).

Nesta seção serão apresentados os principais algoritmos de alinhamento de sequências de nucleotídeos e aminoácidos de bioinformática assim como exemplos de suas utilizações e os *softwares* existentes no mercado que usam estes algoritmos. Serão apresentados também os tipos de alinhamento de sequência mais utilizados pelos usuários.

2.5.1 Métodos de comparação

Uma das formas comparações entre sequências é a realização através de interfaces gráficas amigáveis (GUI - *Graphical User Interface*) consultando os bancos de dados existentes. Comparando sequências, o usuário pode realizar descobertas, são estas: antepassados de organismos, árvores filogenéticas, estruturas de proteínas, funções de proteínas e estruturas genéticas (CELESTI *et al.*, 2019).

Celesti *et al.* (2019) afirma que existe outra forma de fazer comparação de sequências por meio de linhas de comando, onde são executados sem a interface gráfica dos sistemas

[3] A replicação é o processo de duplicação de uma molécula de DNA de dupla cadeia (hélice)

[4] Árvores filogenéticas são árvores genealógicas que são construídas a partir de informações obtidas na comparação das sequências de aminoácidos de uma proteína.

operacionais, geralmente neste caso é utilizado o Linux[5].

Os principais bancos de dados que contém informações de sequenciamentos são: *Genbank*, um banco de dados público de sequência de nucleotídeos e aminoácidos. Este banco de dados foi produzido e está disponível no *National Center for Biotechnology Information* (NCBI); *Protein Information Resource (PIR-International)*, banco de dados de sequências de proteínas, criado pelo *National Biomedical Research Foundation* (NBFR) em 1960 e o *Swiss-Prot*, banco de dados de proteínas mantido por *Swiss Institute of Bioinformatics* (SIB) e Uniprot (WATSON, 2009).

Além dos dados da sequência, os bancos de dados concentram informações como: nome e a classificação da proteína e do organismo onde são encontradas; referência à literatura principal, incluindo informações sobre a determinação da sequência; características funcionais e gerais da proteína; locais de interesse biológico[6] dentro da sequência; entre outras (NORMAND *et al.*, 2018).

Existem dois conceitos comuns de bancos de dados biológicos: bancos de dados primários e bancos de dados secundários. Os bancos de dados primários carregam novos dados explorados em experimentos e atualizam as entradas para garantir a qualidade dos dados, geralmente estes bancos contêm apenas um tipo de dado específico. Diferente dos bancos de dados primários, os secundários usam outros bancos de dados como fonte de informações, portanto, obtêm seus dados solicitando outros bancos de dados, geralmente processam ou analisam os dados correspondentes à solicitação correspondente para obter novos resultados (SZKLARCZYK *et al.*, 2011).

2.5.2 Alinhamento de sequências

O alinhamento de sequências consiste no processo de comparar duas ou mais sequências (nucleotídeos ou aminoácidos) de forma a se observar o seu nível de similaridade. Esta técnica de comparação de sequências é implementada segundo um conceito de desenvolvimento de programas conhecido como algoritmo guloso[7], um dos pilares da bioinformática (GU; BOURNE, 2009).

De acordo com (LESK, 2019), exitem dois tipos de alinhamentos:

- Alinhamento Global;
- Alinhamento Local.

Para cada tipo de alinhamento existem três métodos de pontuação: identidade, similaridade e homologia. A identidade refere-se a presença do mesmo nucleotídeo ou aminoácido

[5] Sistema Operacional lançado na década de 90 por Linus Torvalds, tendo em toda sua trajetória de vida várias versões e o diferencial por ser gratuíto

[6] Locais onde podem conter sequências de nucleotídeos ou aminoácidos já conecidos, onde o pesquisador deseja verificar sua ocorrência em outro organismo.

[7] Algoritmo guloso é ideal para situações de otimização. A cada decisão irá escolher a alternativa mais promissora.

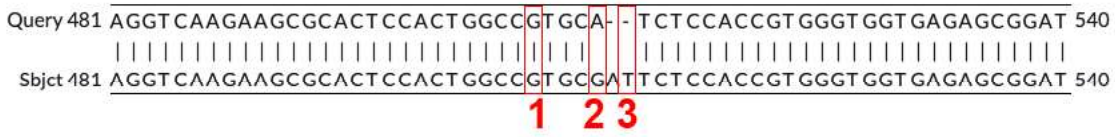
na mesma posição em duas sequências alinhadas. Similaridade é dada pelo alinhamento que tem maior pontuação entre todos os alinhamentos possíveis e homologia dividem a mesma ancestraliade com significado evolutivo (SOLOMON *et al.*, 2016).

A identificação de uma sequência homóloga, se dá por meio da utilização de matrizes de pontuação. *Percent Accepted Mutation* (PAM) e *Blocks Substitution Matrices* (BLOSUM), são dois tipos de matrizes comumente utilizadas. As matrizes PAM listam a probabilidade de alteração de um aminoácido para outro nas sequencias homólogas de proteinas durante a evolução, concentrando-se no rastreamento das origens evolutivas das proteínas. A Matriz BLOSUM baseia-se na substituição de pontuações encontradas em vários períodos evolutivos e são utilizadas para encontrar domínios conservados. A matriz PAM verifica mutações em sequência de proteínas evolutivamente relacionadas e a matriz BLOSSUM em proteínas evolutivamente divergentes (MOUNT, 2008).

2.5.2.1 Alinhamento global

Segundo Hashemifar *et al.* (2016), o alinhamento global é aquele que se estende por toda a sequência, buscando o alinhamento em toda a extensão de duas ou mais sequências que serão comparadas. Este algoritmo foi desenvolvido por Needleman e Wunsch (1970) (algoritmo Needleman-Wunsch) e faz alinhamentos maximizando o número de alinhamentos individuais que são iguais (match) representado na Figura 5 pelo número 1 e minimizando os espaçamentos (GAP) representado pelo número 3. A ausência do *pipe* (| ou barra que indica uma similaridade) significa que não foi alinhado e quando não houver similaridade (*mismatch* representado pelo número 2) entre a *query* e o Sbjct apresentando um GAP.

Figura 5 – Alinhamento Global



Fonte: Desenvolvido pelo autor.

Jararweh *et al.* (2019), desenvolveram um método para otimizar a performance do algoritmo Needleman-Wunsch usando paralelização[8] e técnicas de vetorização. Existem dois métodos clássicos de paralelização, um deles identifica o paralelismo através da compreensão da formulação matemática e o outro método, por meio da submissão de um código em um compilador paralelizador que extrai, automaticamente, oportunidades de paralelismo.

[8] Computação paralela é uma forma de computação em que vários cálculos são realizados ao mesmo tempo, operando sob o princípio de que grandes problemas geralmente podem ser divididos em problemas menores, que então são resolvidos concorrentemente (em paralelo)

2.5.2.2 Alinhamento local

O algoritmo de alinhamento local, conhecido também como Smith-Waterman foi proposto por Smith, Waterman *et al.* (1981). Este algoritmo busca regiões similares, não importando as sequências adjacentes a estas regiões e faz uma procura por regiões com semelhança local não considerando a sequência em todo o seu comprimento (NOTREDAME; HIGGINS; HERINGA, 2000).

Este alinhamento é largamente utilizado para identificar trechos altamente conservados entre dois genomas e utiliza uma pontuação máxima entre qualquer par de sequências.

2.5.3 *Softwares* de alinhamento

Softwares de alinhamento têm a função de verificar sequências de nucleotídeos ou aminoácidos submetidas pelo usuário e encontrar possíveis similaridades com dados já depositados em bancos de dados. Os alinhamentos de pares (*pairwise*[9]), são obtidos por meio dos resíduos entre as sequências, que podem assim fornecer pistas para determinação da possível função das proteínas. Estes *softwares* ajudam o pesquisador a encontrar semelhanças (regiões conservadas), diferenças e mutações entre sequências já conhecidas, depositadas nos bancos de dados. Portanto, a utilização destes programas que realizam este tipo de análise, trazem para o pesquisador uma velocidade maior na obtenção dos dados pesquisados (KATOH; TOH, 2008).

Nesta Seção serão apresentados os *softwares* de alinhamento utilizados para identificação de novas sequências de nucleotídeos e aminoácidos, proteínas e cadeias de DNA e RNA.

2.5.3.1 ClustalW / ClustalX

De acordo com Thompson, Higgins e Gibson (1994), o ClustalW é um programa de alinhamento de múltiplas sequências, podendo ser grupos de nucleotídeos ou sequências de aminoácidos. Os alinhamentos múltiplos são usados para caracterizar famílias de proteínas, identificar homologias entre famílias de sequências conhecidas, para ajudar a prever estruturas secundárias e terciárias de novas sequências, sugerir primers para Polymerase Chain Reaction (PCR), entre outros. Esses resultados obtidos com os alinhamentos de sequências são essenciais para a biologia molecular, biologia evolutica, bioquímica e outras áreas da genética, pois com eles pode-se estabelecer identidades entre sequências, permitindo também a dedução de funções de proteínas baseadas na similaridade, pode-se identificar domínios proteicos conservados e o estudo da evolução de proteínas.

Para os alinhamentos múltiplos serem realizados, podem ser necessárias várias horas de processamento em *clusters* (OLIVER *et al.*, 2005). Os alinhamentos progressivos utilizam métodos heurísticos para fazer os múltiplos alinhamentos os quais são divididos em três passos: o

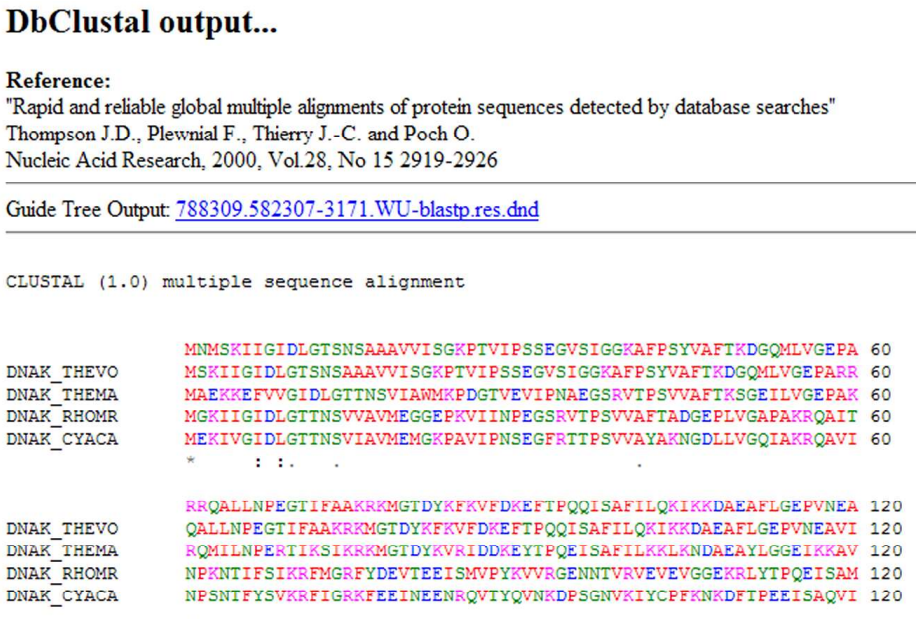
[9] Alinhamento Pairwise, é um alinhamento ocorrido entre pares de resíduos, o qual recebe uma pontuação ótima (i.e. pares com máxima semelhança).

primeiro é verificar a distância entre cada par na sequência, o segundo passo é construir a árvore filogenética baseada na distância da matriz e, por último, o alinhamento de pares de vários perfis é realizado seguindo a ordem de ramificação na árvore filogenética (THOMPSON; HIGGINS; GIBSON, 1994).

Hung *et al.* (2015) afirma que uma técnica utilizando Graphics Processing Unit (GPU), tem um aumento de performance de até 33 vezes no processamento paralelo de sequências comparado com o processamento normal. A GPU é uma arquitetura importante para pesquisas com programação paralela, em virtude do aumento de performance os cientistas podem trabalhar com mais amostragens e assim extrair mais dados simultaneamente.

O Clustal possui duas versões ativas, o ClustalW (linhas de comando) e o ClustalX (interface gráfica). Ambos podem ser executados nos sistemas operacionais Windows, Linux e MacOS[10] (CLUSTALW/CLUSTALX, 2020). A Figura 6 exemplifica a interface de resultados de uma análise de aminoácidos realizada pela versão ClustalX, tendo sido realizado o alinhamento de múltiplas sequências.

Figura 6 – ClustalX - Interface de Resultados



Fonte: ClustalW/ClustalX (2020).

2.5.3.2 Clustal Omega

De acordo com Sievers e Higgins (2018), o Clustal Omega é muito utilizado para sequenciamentos múltiplos e os alinhamentos são baseados em comparações ou previsões de estruturas de proteínas.

[10] Sistema Operacional proprietário da empresa Apple, desenvolvido inicialmente em 1989, baseado também em um *kernel Unix* semelhante ao Linux

O Clustal Omega faz alinhamentos com maior velocidade e retorna resultados mais precisos, trabalhando com sequências maiores, superando outros programas com relação ao tempo de execução e a quantidade de memória necessária para fazer os alinhamentos (SIEVERS *et al.*, 2013).

Os programas da família Clustal são gratuitos e seus executáveis e documentações estão disponíveis para *download* no *website* do fabricante (www.clustal.org).

2.5.3.3 *Molecular Evolutionary Genetics Analysis*

O *Molecular Evolutionary Genetics Analysis* (MEGA) é um software que foi desenvolvido para facilitar a obtenção de informações, mineração de dados e análise de sequências de DNA (KUMAR; TAMURA; NEI, 2004). Criado na década de 90, tendo oito versões para o sistema operacional *Windows* e atualmente, em sua décima versão (MEGA10) é executado nativamente em Linux e MacOS (KUMAR *et al.*, 2018) (MEGA10, 2020).

O MEGA, em todas as suas versões, implementa métodos para análise filogénica, infere árvores evolutivas e estima distâncias evolutivas entre organismos (KUMAR *et al.*, 2018). A Figura 7 exemplifica a interface de resultados de análise exibida pelo programa MEGA10 (disponível em https://www.megasoftware.net).

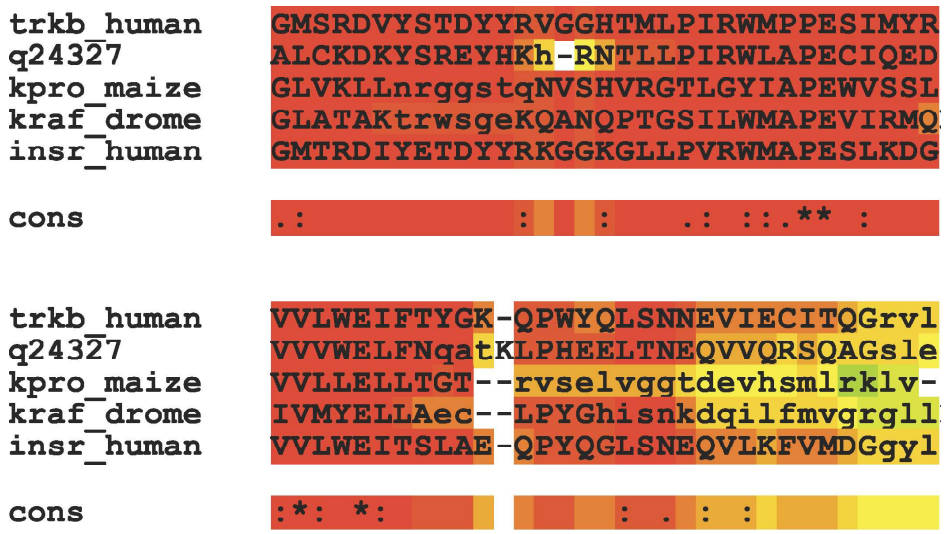
Figura 7 – MEGA10 - Interface de Resultados

Fonte: MEGA10 (2020).

2.5.3.4 *Multiple Sequence Comparison by Log-Expectation*

O *Multiple Sequence Comparison by Log-Expectation* (MUSCLE[11]) comparado aos *softwares* da família Clustal, faz alinhamentos de sequências com mais rapidez e exatidão. O *software* faz alinhamentos progressivos, podendo ser alinhados cinco mil sequências com comprimento médio de 350 pares de base em um curto espaço de tempo, sendo que estas execuções são feitas em computadores pessoais (com processador de dois a quatro núcleos físicos e oito gigas de memória executando o sistema operacional Linux) (EDGAR, 2004). A Figura 8 apresenta a interface de resultados do alinhamento de proteínas.

Figura 8 – MUSCLE - Interface de Resultados



Fonte: Adaptado de Notredame (2007).

De acordo com Pais *et al.* (2014), os erros nos alinhamentos de múltiplas sequências progressivas são comuns e exigem um pós-processamento para correções desses alinhamentos. No pós-processamento, o usuário pode realizar refinamentos iterativos baseados em uma técnica chamada particionamento restrito. O refinamento interativo consiste em refinar alinhamentos iniciais produzidos por um algoritmo até uma condição de parada.

2.5.3.5 *Multiple Alignment Using Fast Fourier Transform*

Outro software que trabalha com alinhamento múltiplo de sequências é o *Multiple Alignment Using Fast Fourier Transform* (MAFFT[12]), feito para Sistemas Operacionais baseados em Unix[13], usado em inferência filogenética, detecção de região conservada, previsão da estrutura de RNAs não codificantes (ncRNAs) e proteínas (KATOH; TOH, 2008).

O MAFFT permite o cálculo paralelo de um grande número de sequências, possuindo maior precisão que outros métodos utilizados para grandes quantidades de dados, mas este

[11] disponível em https://www.drive5.com/muscle
[12] disponível em https://mafft.cbrc.jp/alignment/software
[13] Unix - Sistema Operacional criado na década de 60, o qual os Sistemas atuais se baseiam: Linux, MacOSX, FreeBSD e Solaris

método tornou-se impraticável, devido à necessidade de grandes recursos computacionais (ROZEWICKI *et al.*, 2019).

2.5.3.6 *T-Coffee*

Segundo Tommaso *et al.* (2011), o T-Coffee[14] é um *software* de alinhamento múltiplo de sequências que usa uma abordagem progressiva, gerando uma biblioteca de alinhamentos em pares. Este *software* faz uso de algoritmos para alinhar sequências ou combinar a saída dos seus métodos de alinhamento favoritos (Clustal, Mafft, Muscle).

Notredame, Higgins e Heringa (2000) identificam que o alinhamento de três ou mais sequências de nucleotídeos e aminoácidos é uma das tarefas mais comuns em bioinformática. Os alinhamentos múltiplos são essenciais para análise de proteínas, modelagem por homologia e reconstrução filogenética. O alinhamento múltiplo progressivo tem vantagens na velocidade do alinhamento, na simplicidade e menor tendência de cometer erros no alinhamento, sendo uma caracteristica do *software* T-Coffee.

2.5.3.7 *BLAST*

De acordo com Altschul (1997), *Basic Local Alignment Search Tool* (BLAST[15]) é uma ferramenta amplamente utilizada que busca as similaridades de nucleotídeos e aminoácidos em bases de dados. São feitas comparações em sequências biológicas primárias, como aminoácidos nas proteínas e nucleotídeos em sequências de DNA.

O BLAST é uma ferramenta que pode ser utilizada tanto on-line numa interface *web* quanto *stand-alone* no computador (JOHNSON *et al.*, 2008). Esta ferramenta tem como resultado sequências com alta significância estatística e similaridade à sequência consultada. Os resultados obtidos apesar de sua similaridade não são homólogos. No portal NCBI, existem variações disponíveis do BLAST, são elas BLASTN, BLASTX, BLASTP, TBLASTN e TBLASTX (WATSON *et al.*, 2015).

O BLASTP faz comparações de sequências de aminoácidos no banco de dados de proteínas, já o BLASTN compara sequências de ácidos nucléicos no banco de dados de ácidos nucléicos e finalmente, o BLASTX traduz a sequência de ácidos nucléicos em aminoácidos e compara na base de dados de proteínas. O TBLASTN, faz o processo reverso do BLASTX, convertendo uma sequência de aminoácidos em uma sequência de nucleotídeos e compara com os dados depositados em bancos de dados. Por último, o TBLASTX compara as traduções de seis *frames* de uma sequência de nucleotídeos com as traduções de seis *frames* de um banco de dados de sequência de nucleotídeos. A escolha de qual BLAST usar depende da pergunta biológica e do resultado que se espera obter (ALTSCHUL, 1997).

[14] disponível em http://tcoffee.crg.cat
[15] disponível em https://blast.ncbi.nlm.nih.gov/Blast.cgi

A Tabela 1 exibe as diferenças dos algoritmos *BLAST*. A coluna algoritmo identifica qual o tipo será executado, a coluna *query* apresenta qual o tipo do dado que o usuário deverá inserir para a busca. *Database* é a coluna que mostra qual o banco de dados que será pesquisado.

Tabela 1 – Variações do *BLAST*

Algoritmo	*Query*	*Database*	Descrição
BLASTN	nt	nt	Compara a sequência de nucleotídeos de entrada contra um banco de dados de sequências de nucleotídeos.
BLASTX	nt*	a.a.	Compara a sequência de nucleotídeos de entrada traduzida para todas as sequências de leitura possíveis contra um banco de dados de sequências de proteínas. É o algoritmo mais utilizado em grandes projetos de sequenciamento, pois permite identificar possíveis proteínas a partir de uma sequência de nucleotídeos desconhecida.
BLASTP	a.a.	a.a.	Compara a sequência de aminoácido de entrada (*query*) contra um banco de dados de sequências de proteínas (*subject*).
TBLASTN	a.a.	nt*	Compara a sequência de aminoácido de entrada contra um banco de dados de sequências de nucleotídeos traduzidas para todas as sequências de leitura possíveis.
TBLASTX	nt*	nt*	Compara as seis sequências de leitura possíveis de um nucleotídeo contra um banco de dados de nucleotídeos traduzidos para todas as sequências de leitura possíveis.

Fonte: Adaptado de https://www.ncbi.nlm.nih.gov/
a.a. = aminoácido, nt = nucleotídeo, * = traduzido para todas as sequências de leituras possíveis.

De forma bastante resumida, o algoritmo BLAST pode ser dividido em três estágios básicos:

1. Compilação de uma lista de "palavras" de alta pontuação;
2. Procura destas palavras no banco de dados;
3. Extensão de alinhamentos a partir das palavras encontradas (ALTSCHUL, 1997).

No primeiro estágio é gerada uma lista de todas as “palavras” encontradas na sequência que está sendo buscada. “Palavra”, no caso do *BLAST*, é um trecho de comprimento definido da sequência. Os tamanhos-padrão de palavras são 3 aminoácidos para sequências protéicas e 11 nucleotídeos para sequências nucléicas (CLAVERIE; NOTREDAME, 2006).

No segundo estágio "Palavras"idênticas às presentes na lista são identificadas nas sequências do banco de dados. Como o banco de dados teve sua lista de palavras pré-compilada e

indexada no momento em que foi criado, não no instante em que se executa a busca, este passo é rápido, demorando de 30 à 60 segundos para execução total (CLAVERIE; NOTREDAME, 2006).

E no terceiro estágio, se o programa identifica um alinhamento com alto grau de continuidade, é produzido um alinhamento de comparação, realizada uma extensão deste mesmo alinhamento, a partir de adaptação do algoritmo de Smith-Waterman e os resultados estatisticamente significativos são demonstrados graficamente para o usuário (CLAVERIE; NOTREDAME, 2006).

Os *softwares* apresentatos executam tarefas importantes aos pesquisadores, facilitando a obtenção de dados tratados para realisar análises. Tendo em vista o cenário de utilização de programas de computadores para obtenção de dados e análise de dados obtidos faz-se necessário a produção de novos softwares que executem e analisem mais rapidamente os dados.

3 JUSTIFICATIVA

O desenvolvimento de uma ferramenta que agregue diversas funcionalidades baseadas em bioinformática pode facilitar e diminuir o tempo que o usuário despenderá para o processamento dos dados coletados em laboratório, evitando a utilização de ferramentas distribuídas de forma esparsa e organizando todos os dados necessários em um só ambiente.

4 MOTIVAÇÃO

A utilização de uma ferramenta agregadora de funcionalidades de bioinformática auxilia no trabalho do usuário.

5 OBJETIVOS

5.1 OBJETIVO GERAL

Desenvolver uma aplicação web agregadora de funcionalidades de bioinformática comuns ao cotidiano do usuário, distribuída gratuitamente e de código aberto.

5.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos são:

- Verificar junto à usuários, quais ferramentas de bioinformática são relevantes nas suas rotinas de trabalho diário;
- Desenvolver a ferramenta proposta;
- Exibir a função das proteínas buscadas;
- Exibir os artigos e textos relacionados aos dados submetidos;
- Realizar testes com usuários sobre a utilidade da ferramenta;
- Analisar os dados dos testes e com isso, aprimorar a ferramenta desenvolvida.

6 MATERIAL E MÉTODOS

A partir dos objetivos traçados na Seção 5, o desenvolvimento do *software* foi realizado na Universidade de Ribeirão Preto – UNAERP e Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, campus São João da Boa Vista (IFSP-SBV).

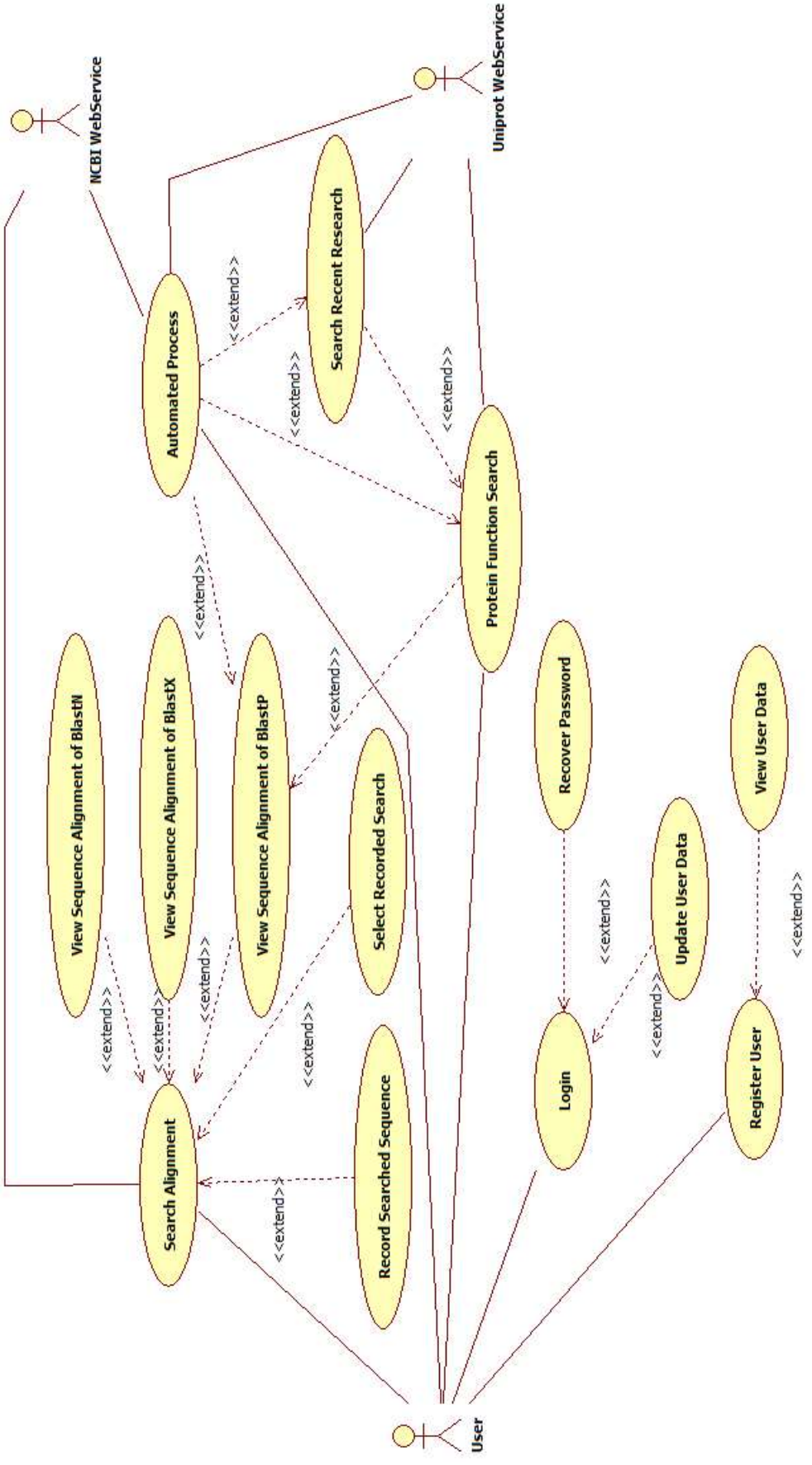
Na primeira etapa do estudo focou-se na identificação do sistema operacional que seria utilizado para o desenvolvimento, nas linguagens de programação, na performance que o servidor teria com o processamento e nos resultados obtidos.

O sistema operacional selecionado para o desenvolvimento do projeto foi o Linux, devido a versatilidade com hospedagem de sistemas *web* e o *software* Apache como provedor do serviço *web* à ferramenta. Como requisitos de *hardware*, foi escolhido um processador Intel Xeon com dois núcleos de três Gigahertz (GHZ), três Gigabytes (GB) de memória *Read Access Memory* (RAM) e cinco GB de armazenamento para *download* de arquivos no formato *Extensible Markup Language* (XML) dos servidores do NCBI e Uniprot.

Na Etapa de desenvolvimento da ferramenta, implementou-se uma ferramenta que faz consultas ao BLAST. Estas consultas são realizadas mediante a inserção de uma *query*[1] no campo do formulário, desenvolvido com a linguagem de marcação *HyperText Marckup Language* (HTML) e processados com *Hypertext Preprocessor* (PHP). Os dados inseridos no cadastro de usuários dentro da ferramenta serão persistidos utilizando o Sistema de Gerenciamento de Banco de Dados (SGBD) onde foi escolhido o MySQL, que utiliza *Structured Query Language* (SQL) como linguagem. Ainda na primeira etapa do estudo fez-se necessário por meio dos requisitos do software, desenvolver um diagrama de caso de uso utilizando a *Unified Modeling Language* (UML), Figura 9, o qual exibe as funcionalidades do *software* por completo.

[1] Dados que o usuário deverá inserir no *software* para que sejam consultados.

Figura 9 – Diagrama de Caso de Uso



Fonte: Desenvolvido pelo autor.

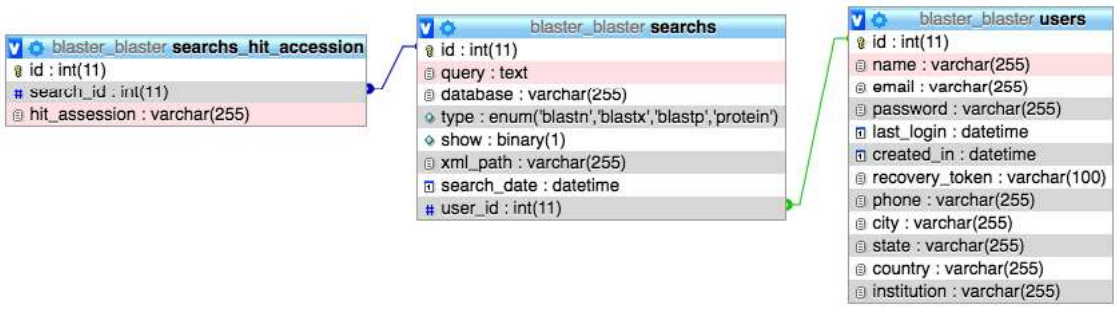
Na Figura 9, são identificados três atores no funcionamento deste *software*, o usuário, o servidor do NCBI e o servidor do UniProt identificados no diagrama pelo desenho de um ator, sendo o usuário a pessoa que irá utilizar o *software*, o servidor do NCBI a fonte onde será consultada a sequência submetida pelo usuário e o servidor do UniProt que será o provedor das informações das funções das proteínas juntamente com artigos relacionados aos alvos encontrados.

Ainda na Figura 9, foram identificados dezesseis casos de uso representados pelos círculos os quais são cada uma das funcionalidades da ferramenta e serão descritas na Seção 6.2.

6.1 MODELO DO BANCO DE DADOS

O banco de dados do BLASTER, foi modelado a partir das necessidades encontradas (Figura 10). Foi necesário uma tabela para persistência dos dados dos usuários do sistema e outras duas tabelas para a armazenamento dos dados pesquisados pelo usuário.

Figura 10 – Modelo do Banco de Dados



Fonte: Desenvolvido pelo autor.

Neste modelo apresentado na Figura 10, podem ser vistas três tabelas: *user*, *searchs* e *searchs_hit_accession*. A tabela *user* tem a função de armazenar todos os dados dos usuários cadastrados. A tabela *searchs* tem a função do armazenamento das pesquisas feitas pelo usuário, onde são armazenados a *query* pesquisada, o banco de dados selecionado (representado pelo campo *database*), o tipo do algoritmo utilizado, o campo *xml_path* o qual armazena o caminho do arquivo XML com os resultados da pesquisa, o campo *search_date* para a data e hora da pesquisa e por último o código do usuário para identificar a qual usuário pertence a busca. Por último a tabela *searchs_hit_accession* foi criada com o intuíto de armazenar os dados relacionados as pesquisas realizadas, portanto é uma tabela de suma importância, pois o usuário cadastrado poderá consultar suas pesquisas já realizadas, sem a necessidade de realizar uma nova busca no BLASTER.

6.2 CASOS DE USO

Nesta Seção serão descritos casos de uso do *software* desenvolvido e suas funcionalidades. As Tabelas nas seções subsequentes contém dois cenários: Principal e Alternativo, onde Cenário Principal corresponde ao fluxo de funcionalidades que iniciam e terminam sem nenhum tipo de contratempos, executando em sua operação normal. O Cenário Alternativo corresponde a ações que podem ocorrer durante sua execução, ações estas que podem interferir o fluxo normal da funcionalidade.

6.2.1 *Register User*

A funcionalidade *Register User* refere-se ao cadastramento de usuários no sistema, esta ação dentro do sistema é através do ítem *Sign up* na barra superior. Para o cadastramento são necessários os seguintes dados: nome, e-mail, senha, cidade, estado, país, instituição e telefone. Todos os campos são obrigatórios com exceção do telefone. Novos cadastros com e-mails já cadastrados no sistema não serão permitidos. A Tabela 2 apresenta o cenário principal e os alternativos deste caso de uso.

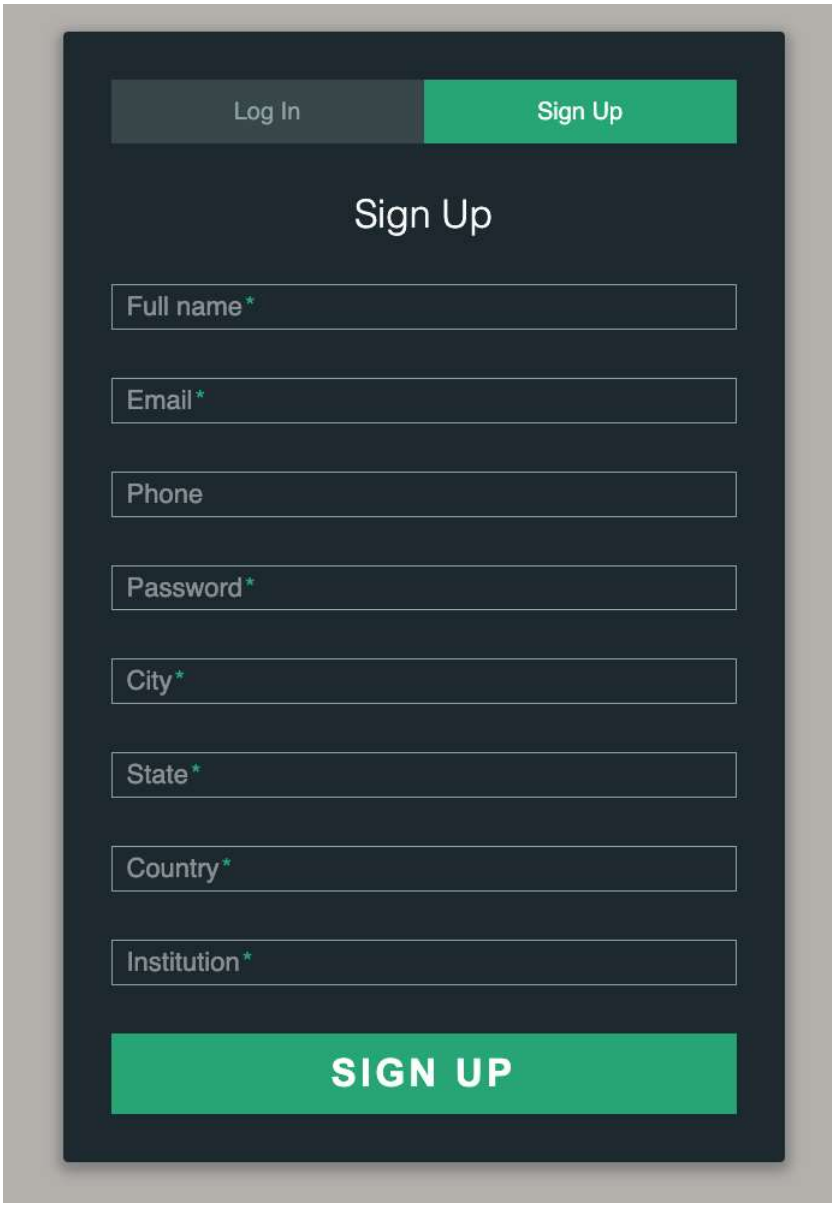
Tabela 2 – Caso de uso: *Register User*

Cenário Principal:	Usuário informa todos os dados solicitados, sistema realiza o cadastro com sucesso, um e-mail de ativação do cadastro é enviado ao usuário para validar o cadastro.
Cenário Alternativo:	1. O usuário não preenche todos os dados. 2. O e-mail informado está com formato inválido 3. O e-mail informado já consta no cadastro 4. Erro de conexão por falta de acesso à Internet

Fonte: Desenvolvida pelo Autor.

A Figura 11 refere-se a tela de registro de usuário para acesso a sistema. Estes campos serão utilizados em futuras funcionalidades para estatística de acessos.

Figura 11 – *Register User - Sign up*



Fonte: Desenvolvido pelo Autor.

6.2.2 *View User Data*

View User Data é uma funcionalidade com a ação de exibir os dados cadastrados do usuário. O campo de telefone pode estar vazio, pois não é um campo de preenchimento obrigatório no cadastro. A Tabela 3 apresenta o cenário principal e os alternativos deste caso de uso.

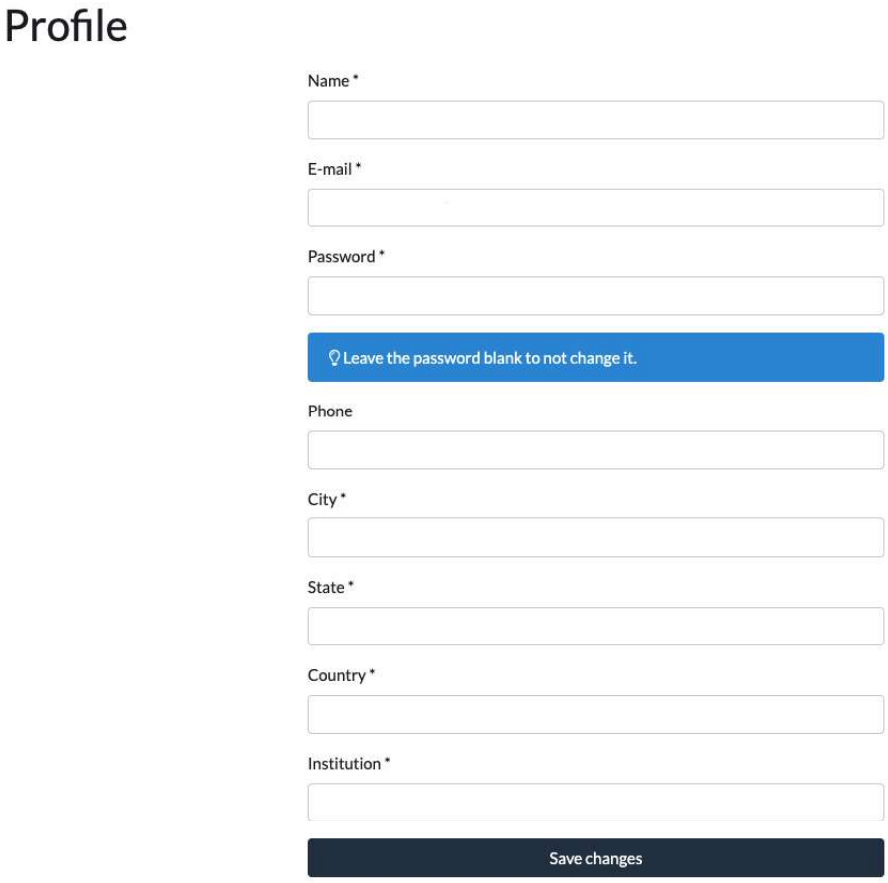
Tabela 3 – Caso de uso: *View User Data*

Cenário Principal:	Usuário visualiza os seus dados cadastrados no banco de dados.
Cenário Alternativo:	1. O usuário visualiza seus dados cadastrados. 2. O campo de telefone não pode ser visualizado. 3. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

A Figura 12 representa a tela de visualização e alteração dos dados (Seção 6.2.3) do usuário cadastrado. Esta tela esta disponível para o usuário dentro do ambiente controlado do sistema através do menu *profile*.

Figura 12 – *View User Data* - *Profile*



Fonte: Desenvolvido pelo autor.

6.2.3 *Update User Data*

Update User Data tem a função de atualizar os dados do usuário que estão cadastrados no banco de dados. O usuário cadastrado não poderá alterar o nome e e-mail do cadastrado, serão alterados somente a senha, cidade, estado, país e telefone. A Tabela 4 apresenta o cenário principal e os alternativos deste caso de uso.

Tabela 4 – Caso de uso: *Update User Data*

Cenário Principal:	O usuário pode altera a senha, cidade, estado, país e telefone do seu cadastro. Salvar os dados no banco de dados.
Cenário Alternativo:	1. O usuário altera e-mail cadastrado. 2. O usuário altera nome cadastrado. 3. O usuário retira a senha do campo senha. 4. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

6.2.4 *Login*

Login refere-se à autenticação do usuário. Os dados para acesso ao sistema são o e-mail e a senha previamente cadastrados. Usuários que não lembrarem a sua senha podem utilizar a funcionalidade de *Recover Password* para recebê-la em seu e-mail (Seção 6.2.5). A Tablea 5 apresenta o cenário principal e os alternativos deste caso de uso.

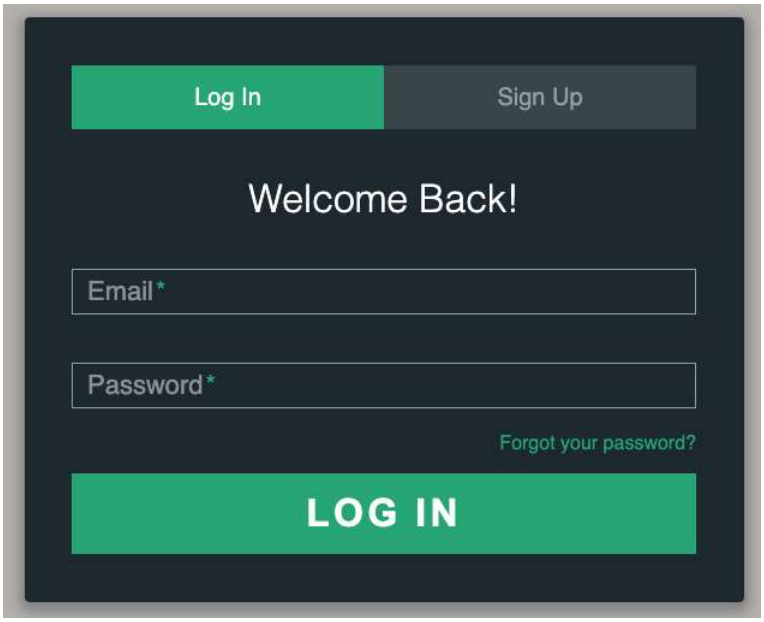
Tabela 5 – Caso de uso: *Login*

Cenário Principal:	O usuário preenche os campos de e-mail e senha e clica no botão *LOG IN*; o sistema realiza a autenticação com sucesso e exibe a tela do *software*, restrita a usuários cadastrados.
Cenário Alternativo:	1. O usuário não preenche todos os dados. 2. O e-mail informado está com formato inválido. 3. O e-mail informado não consta cadastro. 4. A senha informada é inválida. 5. O usuário opta por recuperar senha. 6. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

A Figura 13 refere-se a tela onde o usuário entrará em um ambiente controlado. O usuário deve estar devidamente cadastrado como usuário e será necessário o preenchimento do e-mail e senha para acessar sistema.

Figura 13 – *Login*



Fonte: Desenvolvido pelo Autor.

Neste Ambiente controlado, o usuário poderá ver suas pesquisas já realizadas e salvas através da caixa de seleção *Save to my history* (Seção 6.2.11), como pode ser visto na Figura 17.

6.2.5 *Recover Password*

O *Recover Password* tem a funcionalidade de recuperar a senha do usuário cadastrado no *software*. O usuário deverá fornecer o e-mail para que o sistema consulte a base de dados e envie um e-mail ao usuário com um *link*, o qual poderá criar uma nova senha. Esta funcionalidade está disponível somente para usuários desconectados do sistema. A Tabela 6 apresenta o cenário principal e os alternativos deste caso de uso.

Tabela 6 – Caso de uso: *Recovery Password*

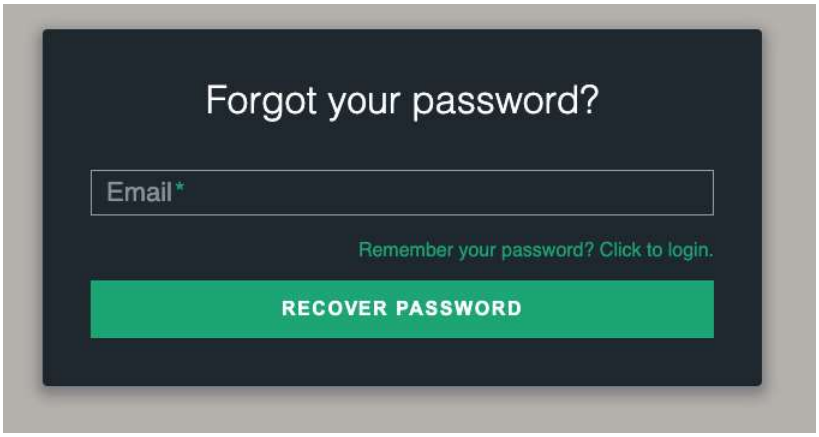
Cenário Principal:	O usuário clica no botão *Recover Password* e preenche o campo e-mail, o sistema verifica o e-mail cadastrado e envia um e-mail ao usuário cadastrado um *link* para criar uma nova senha.
Cenário Alternativo:	1. O usuário não preenche o campo e-mail. 2. O e-mail informado está com formato inválido. 3. O e-mail informado não consta cadastro. 4. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

A Figura 14 refere-se a tela de recuperação do acesso do usuário o sistema. O preenchimento do campo e-mail é obrigatório, para que possa ser enviado um e-mail ao usuário cadastrado o *link* para criar uma nova senha de acesso. Ainda nesta tela, há um *link Remember*

your password? Click to login o qual refere-se a ação de voltar a tela de *login* descrita na Figura 13.

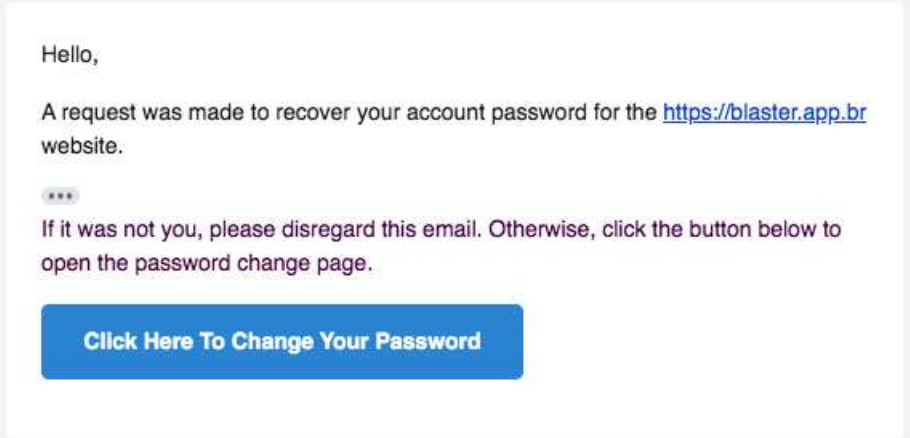
Figura 14 – *Recovery Password*



Fonte: Desenvolvido pelo Autor.

A Figura 15 representa o e-mail que o usuário cadastrado recebe quando usa a funcionalidade *Forgot your password?*. Neste e-mail recebido contém um *link* para recuperação da senha esquecida.

Figura 15 – E-mail de *Recovery Password*

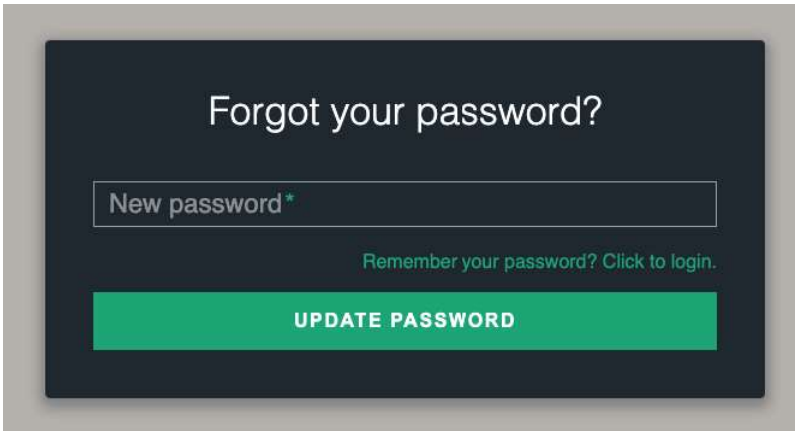


Fonte: Desenvolvido pelo Autor.

A Figura 16 refere-se a tela a qual o *link* enviado por e-mail levará o usuário para a recuperação de senha. Esta tela possibilita ao usuário atualizar a senha esquecida. Esta funcionalidade é executada através de um *token*[2] de acesso, podendo ser executada somente uma vez por e-mail recebido. O *token* é gerado a cada esquecimento de senha.

[2] Conjunto de caracteres gerado aleatóriamente para utilização em sistemas ou dispositivos os quais necessitam segurança dos dados / contra-senha

Figura 16 – *Recovery Password - Update password*



Fonte: Desenvolvido pelo Autor.

6.2.6 *Search Alignment*

A *Search Alignment* tem a função de buscar os dados submetidos pelo usuário no servidor do NCBI, este por sua vez responde os dados no formato XML. Os dados retornados são processados e exibidos em formato de tabela ao usuário.

6.2.6.1 *Search BLASTN*

Nesta Seção será apresentada a funcionalidade *Search BLASTN* exibida na Tabela 7 na qual são apresentados o cenário principal e os cenários alternativos deste caso de uso.

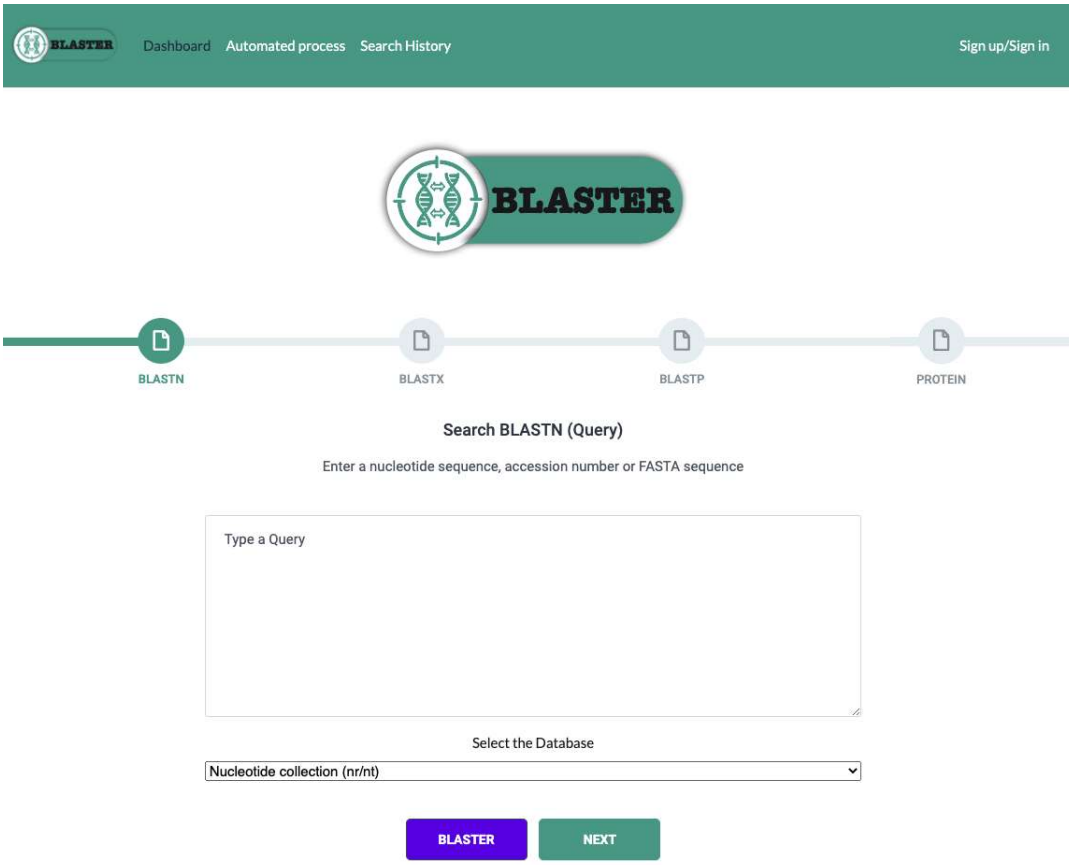
Tabela 7 – Funcionalidade: *Search BLASTN*

Cenário Principal:	O usuário insere no campo *query* uma sequência de nucleotídeos, *Accession Number* ou uma sequência no formato FASTA, seleciona o banco de dados ao qual será realizada a busca e clica no botão BLAST. Após o tempo de processamento, o *software* exibe em formato de tabela os dados encontrados.
Cenário Alternativo:	1. O usuário não preenche o campo *query*. 2. O usuário insere uma sequência com poucos pares de base. 3. Não é encontrada nenhuma sequência similar. 4. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

A Figura 17 é a representação da tela onde o usuário irá submetr os dados. Esta tela é exibida em dois momentos: quando o usuário acessa o endereço onde esta hospedado o *software* diretamente e após o ato de *login* no sistema. A diferença entre os dois acessos é a caixa de seleção *Save to my history*, disponível para os usuários dentro do ambiente controlado.

Figura 17 – Funcionalidade - *Search BLASTN*



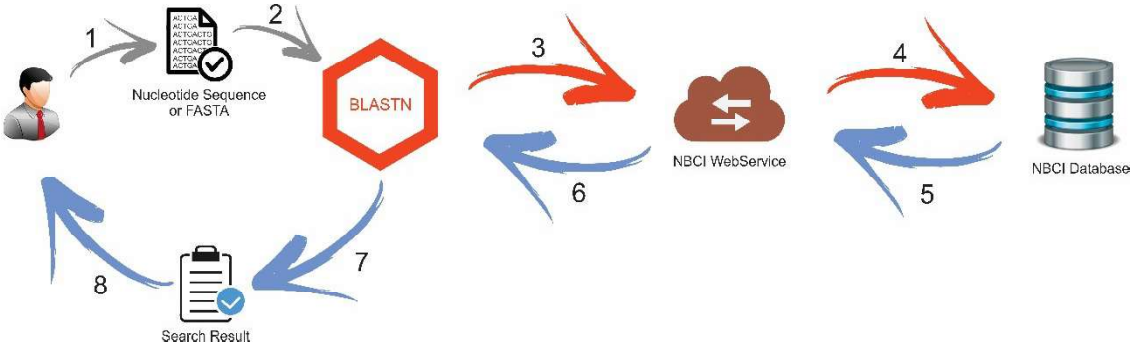
Fonte: Desenvolvido pelo Autor.

A Figura 17 representa a tela do *software* onde o usuário deverá inserir no campo *query* uma sequência de nucleotídeos, um *Accession Number* ou uma sequência FASTA[3] no campo *query* e selecionar a base de dados, para que o *software* realize a busca. O usuário também poderá avançar e ir diretamente ao *Search BLASTX* clicando no botão *NEXT*. Este botão tem dupla função, uma delas, levar o usuário ao próximo passo da *pipeline* e obter as informações preenchidas no campo *query* do *Search BLASTN* preenchendo automaticamente o campo *query* da tela *Search BLASTX*, esta última ação acontece somente se o campo do *Search BLASTN* estiver preenchido. Como cenário alternativo, se o usuário não preencher do campo *query* e tentar executar a ferramenta através do botão *BLAST*, o *software* exibirá uma mensagem de alerta dizendo que o campo *query* não tem nenhum dado.

A Figura 18 exemplifica o funcionamento completo do *Search BLASTN*, estando dividida em oito passos.

[3] Formato para representar sequências de nucleotídeos quanto sequências de peptídeos, no qual os nucleotídeos ou aminoácidos não representados usando códigos de uma única letra (SAUVAGE *et al.*, 2018).

Figura 18 – Fluxo da consulta ao *BLASTN*



Fonte: Desenvolvido pelo Autor.

Os passos numerados de 2 a 8 na Figura 18 são relacionados a cada uma das ações realizadas dentro do *software*, sendo: (1) O usuário seleciona a sequência de nucleotídeos ou FASTA a ser submetido; (2) O usuário insere as informações dentro do campo *query* e seleciona o *database* que será consultado, exemplificado na Figura 18; (3) Após os dados serem submetidos, o *software* BLASTER faz uma consulta ao servidor do NCBI; (4) Por sua vez, o servidor do NBCI consulta sua base de dados; (5) A base de dados retorna ao servidor a resposta à solicitação submetida; (6) O servidor retorna ao BLASTER um arquivo no formato XML; (7) O BLASTER faz a interpretação do arquivo XML e compila uma tabela com os dados de resposta; (8) Por último, os dados são apresentados na tela ao usuário.

O processo descrito nos 8 passos da Figura 18 pode demorar de 25 a 120 segundos em determinados momentos do dia, este tempo é similar ao do próprio portal do NCBI.

Após a consulta realizada (passo 2) e depois de decorrido o tempo necessário para a execução dos processos (passos 3, 4 e 5), é realizado o *download* dos dados no formato XML (passo 6) em segundo plano. Esta requisição é feita utilizando AJAX, sendo realizada de forma assíncrona ao servidor do NCBI.

Neste momento, o arquivo XML contendo os dados extraídos da plataforma BLAST precisa ser interpretado e compilado (passo 7) e apresentado os resultados para o usuário (passo 8), isso se faz com uma requisição AJAX a um arquivo PHP que realiza a leitura do XML e exibe o resultado da busca no formato de tabela dinâmica[4].

6.2.6.2 *Search BLASTX*

A funcionalidade *Search BLASTX* é semelhante à *Search BLASTN* descrito na Seção 6.2.6.1, realizando o mesmo processo de acesso assíncrono feito com AJAX no portal NCBI. O usuário pode utilizar nesta tela uma sequência de nucleotídeos, um *Accession Number* ou uma sequência FASTA e selecionar a base de dados na qual será realizada sua busca. A Tabela 8 apresenta o cenário principal e os alternativos deste caso de uso.

[4] Tabela dinâmica é aquela que os dados podem ser filtrados através de um campo de busca.

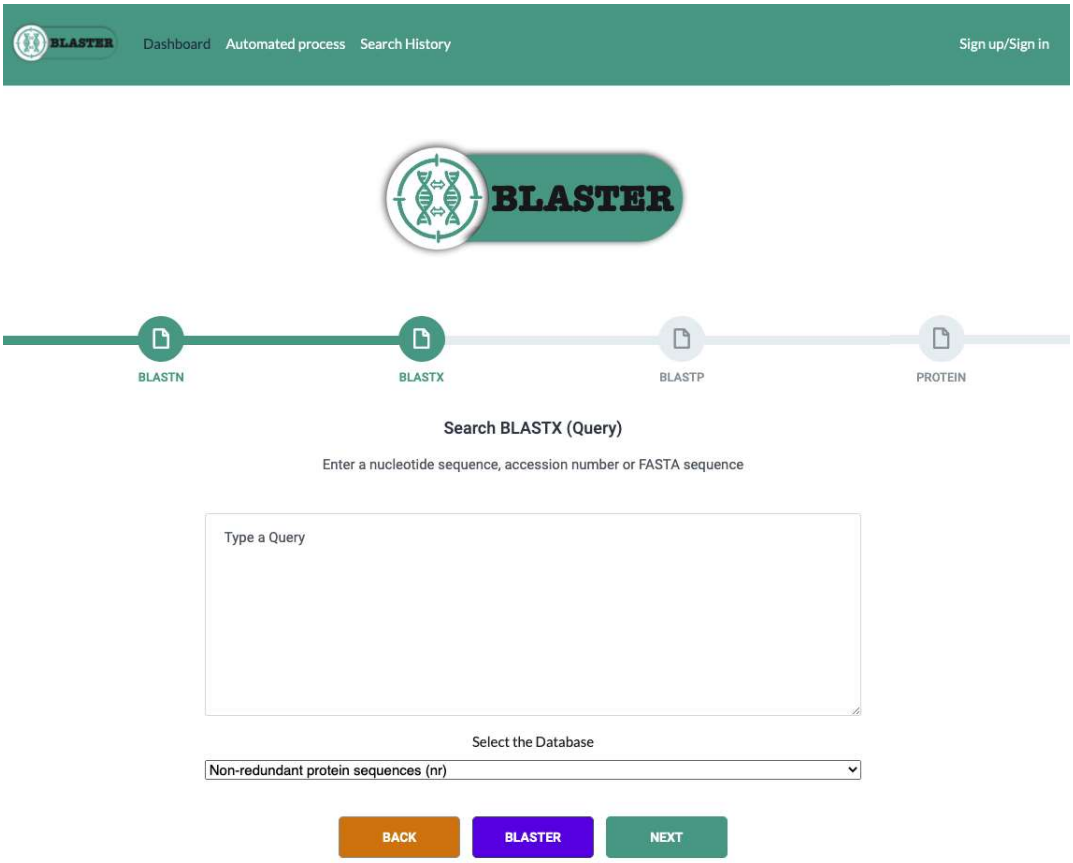
Tabela 8 – Funcionalidade: *Search BLASTX*

Cenário Principal:	O usuário insere no campo *query* um *Accession Number*, uma sequência de nucleotídeos ou uma sequência no formato FASTA; seleciona o banco de dados o qual será realizada a busca e clica no botão BLAST. Após o tempo de processamento, o *software* exibe em formato de tabela os dados encontrados.
Cenário Alternativo:	1. O usuário não preenche o campo *query*. 2. O usuário insere uma sequência com poucos pares de base. 3. Não é encontrada nenhuma sequência similar. 4. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

A Figura 19 representa a tela do *software* em que o usuário deverá inserir um *Accession Number*, uma sequência de nucleotídeos ou a sequência FASTA e selecionar a base de dados, para que o BLASTER realize a busca. Nesta tela ainda, o usuário pode retroceder o passo apertando o botão *BACK* e voltar ao *Search BLASTN* ou apertar o botão *NEXT* e ir ao *Search BLASTP*. Para a Execução do *Search BLASTX* o usu ário deve apertar o botão BLAST. Nesta tela o botão *NEXT* tem somente a função de ir para a próxima tela.

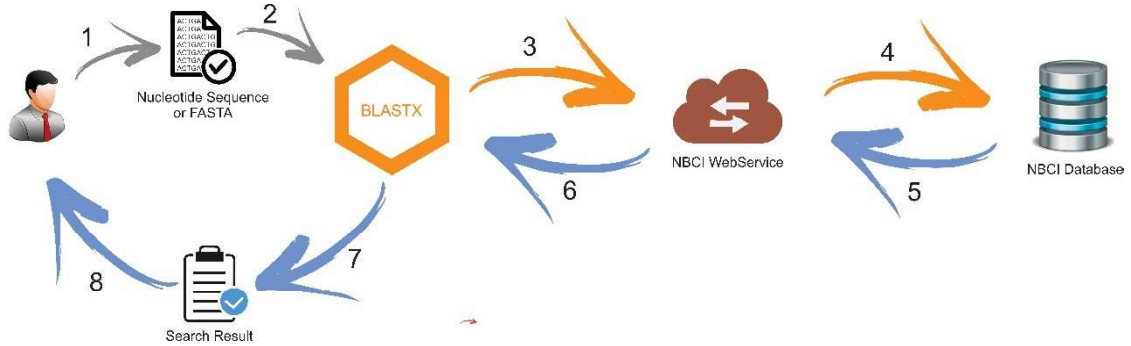
Figura 19 – Funcionalidade - *Search BLASTX*



Fonte: Desenvolvido pelo Autor.

O fluxo do *Search BLASTX* é exemplificado na Figura 20.

Figura 20 – Fluxo da consulta ao *BLASTX*



Fonte: Desenvolvido pelo Autor.

Nesta etapa do *software* descrita na Figura 20, são executadas 8 ações, com as mesmas funcionalidades descritas anteriormente na Seção 6.2.6.1 mudando apenas o algoritmo que será consultado e os passos 6, 7 e 8, que retornarão ao usuário uma informação a mais, o FASTA.

6.2.6.3 *Search BLASTP*

A funcionalidade *Search BLASTP*, busca por proteínas já catalogadas na base de dados do NCBI, se assemelhando as buscas citadas nas Seções 6.2.6.1 e 6.2.6.2, onde o usuário pode selecionar diretamente o *Search BLASTP*, ou passar pelos passos anteriores. Esta funcionalidade diferencia-se do *Search BLASTN* e do *Search BLASTX* pelo fato de não aceitar sequências de nucleotídeos no campo *query*. A Tabela 9 apresenta o cenário principal e os alternativos deste caso de uso.

Tabela 9 – Funcionalidade: *Search BLASTP*

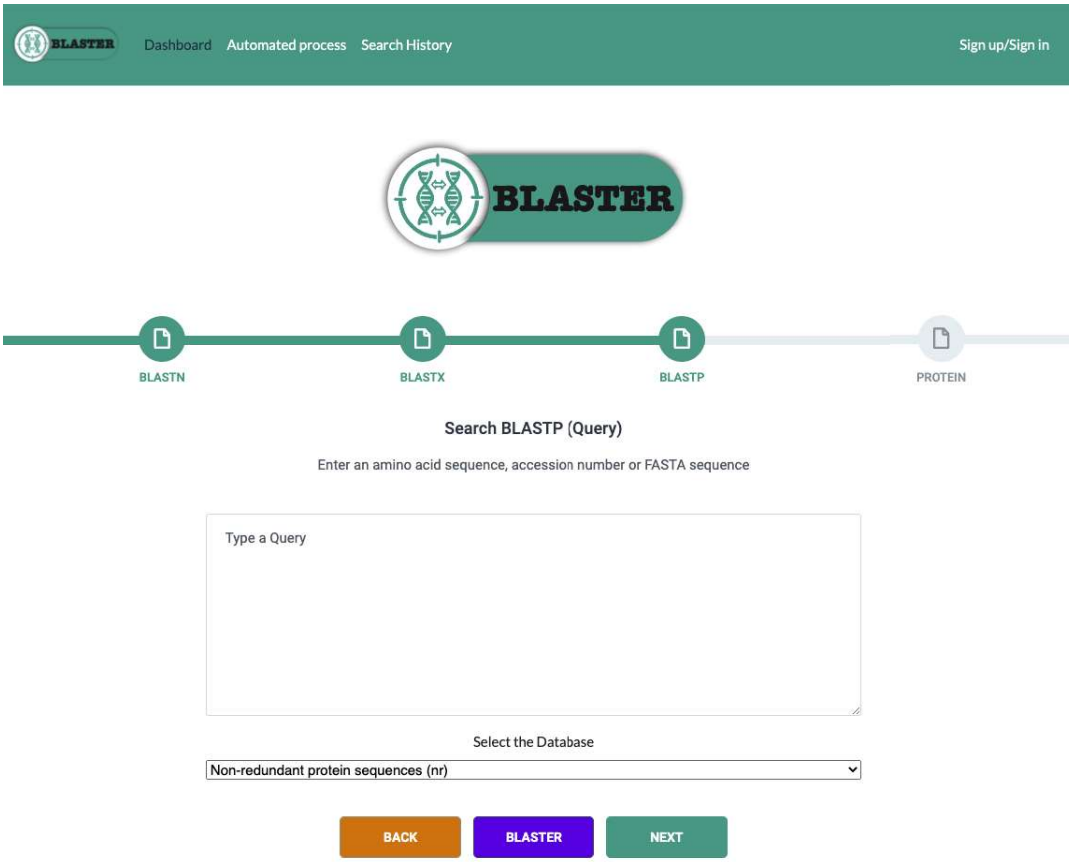
Cenário Principal:	O usuário insereno campo *query* um *Accession Number*, uma sequência de aminoácidos ou uma sequência no formato FASTA, seleciona o banco de dados o qual será realizada a busca e clica no botão *BLAST*. Após o tempo de processamento, o BLASTER exibe em formato de tabela os dados encontrados.
Cenário Alternativo:	1. O usuário não preenche o campo *query*. 2. O usuário insere uma sequência com poucos pares de base. 3. Não é encontrada nenhuma sequência similar. 4. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

A Figura 21 representa a tela em que o usuário deverá inserir uma sequência de aminoácidos, um *Accession Number*, ou a sequência FASTA e selecionar a base de dados, para que o *software* realize a busca. Nesta tela do BLASTER, se desejável ou necessáio, o usuário pode retroceder o passo apertando o botão *BACK* e voltar ao *Search BLASTX*.

O *Search BLASTP* exibe uma tela semelhante à do *Search BLASTN* e *Search BLASTX*, mas tem uma funcionalidade distinta, a de levar a uma tela onde serão exibidas as funções das proteínas encontradas e listadas. Para acessar esta funcionalidade o usuário deve selecionar *hit* para ir ao próximo estágio, este assunto sera tratado na Seção 6.2.12.

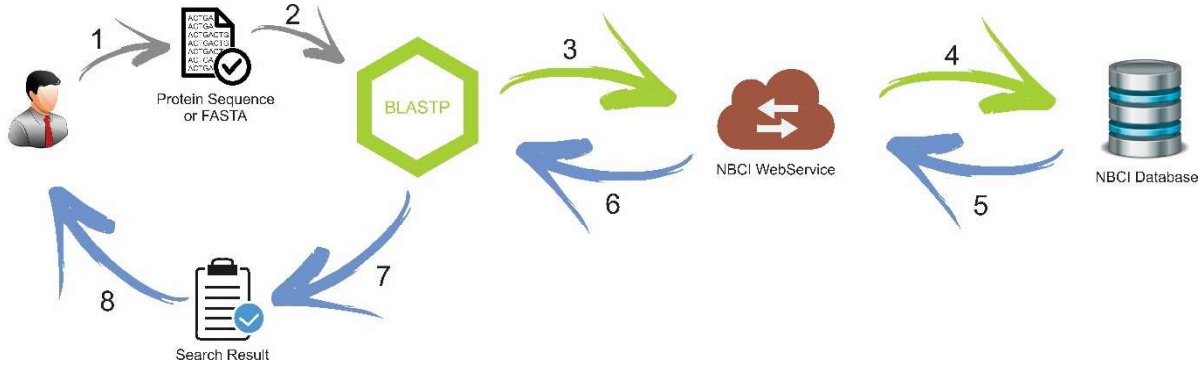
Figura 21 – Funcionalidade - *Search BLASTP*



Fonte: Desenvolvido pelo Autor.

A Figura 22 exemplifica o fluxo de utilização do *software* na funcionalidade *Search BLASTP* e novamente são apresentados 8 passos referentes as ações que serão executadas.

Figura 22 – Fluxo da consulta ao *BLASTP*



Fonte: Desenvolvido pelo Autor.

Os 8 passos da Figura 22, são os mesmos relacionados na Seção 6.2.6.1 com a alteração do algoritmo que nesta etapa utilizará o *BLASTP*. Nos passos 1 e 2 o usuário deverá inserir no campo *query* uma sequência de aminoácidos no formato FASTA ao invés de uma sequência de nucleotídeos. Nos passos 6, 7 e 8 os resultados obtidos retornam informações referentes aos números de identificação da proteínas encontradas. Este número de identificação (*Accession Number*) é utilizado para realizar a busca da função da proteína no portal UniProt.

6.2.7 *View Sequence Alignment of BLASTN*

A funcionalidade *View Sequence Alignment of BLASTN* exibe o alinhamento obtido por meio da consulta ao portal NCBI. Na Tabela 10 é apresentado o cenário principal e os alternativos deste caso de uso.

Tabela 10 – Caso de uso: *View Sequence Alignment of BLASTN*

Cenário Principal:	O usuário clica na linha onde é exibido os *Hits* encontrados na busca e o alinhamento referente ao registro selecionado é exibido.
Cenário Alternativo:	1. O usuário não preenche o campo *query*. 2. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

Esta funcionalidade é descrita na Figura 18, nos passos 7 e 8, sendo a interpretação do arquivo XML recebido do portal NCBI, juntamente com a montagem da tabela dinâmica com os dados retornados e a exibição para o usuário.

6.2.8 *View Sequence Alignment of BLASTX*

View Sequence Alignment of BLASTX tem sua funcionalidade semelhante a função exibida na Seção 6.2.7. Na Tabela 11 é apresentado o cenário principal e os alternativos deste caso de uso.

Tabela 11 – Caso de uso: *View Sequence Alignment of BLASTX*

Cenário Principal:	O usuário seleciona o registro onde são exibidos os *Hits* encontrados na busca e o alinhamento é exibido.
Cenário Alternativo:	1. O usuário não preenche o campo *query*. 2. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

Esta funcionalidade pode ser vista na Figura 20, nos passos 7 e 8, em que o arquivo XML recebido do servidor do NCBI é interpretado, montado a tabela dinâmica com os dados retornados e a exibição para o usuário.

6.2.9 *View Sequence Alignment of BLASTP*

View Sequence Alignment of BLASTP tem sua funcionalidade semelhante a função exibida na Seção 6.2.7 e 6.2.8. Na Tabela 12 é apresentado o cenário principal e os alternativos deste caso de uso.

Tabela 12 – Caso de uso: *View Sequence Alignment of BLASTP*

Cenário Principal:	Ao Usuário clicar na linha onde é exibido os *Hits* encontrados na busca e o alinhamento referente a linha selecionada é exibido.
Cenário Alternativo:	1. O usuário não preenche o campo *query*. 2. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvida pelo Autor.

Esta funcionalidade pode ser vista na Figura 22, nos passos 7 e 8, em que o arquivo XML recebido do servidor do NCBI é interpretado, montado a tabela dinâmica com os dados retornados e a exibição para o usuário.

6.2.10 *Record Searched Sequence*

A funcionalidade *Record Searched Sequence* tem como objetivo a persistência dos dados em um banco de dados. As informações que serão salvas na base de dados são: *id* do usuário, data e horário da utilização do sistema, a *query* submetida a pesquisa e o *Accession Number* do *Hit* escolhido pelo usuário. O usuário tem a opção de não ativar esta função, desmarcando a caixa de seleção *Save to my history*, presente nas telas de *Search BLASTN* (Seção 6.2.6.1), *Search BLASTX* (Seção 6.2.6.2), *Search BLASTP* (Seção 6.2.6.3) e *Search Protein* (Seção 6.2.12). Na Tabela 13 é visto o cenário principal e os alternativos deste caso de uso.

Tabela 13 – Caso de uso: *Record Searched Sequence*

Cenário Principal:	O usuário dentro do sistema controlado do BLASTER, pode salvar suas sequências, sem a necessidade de executar a busca na base de dados do NCBI.
Cenário Alternativo:	1. O usuário não seleciona a caixa *Save to my history* 2. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvido pelo Autor.

6.2.11 *Select Recorded Search*

A funcionalidade *Select Recorded Search* tem como objetivo listar os dados persistidos de um banco de dados. As informações salvas na base de dados são exibidas para o usuário em formato de tabela dentro do ambiente controlado, portanto o usuário deve realizar o *login* no BLASTER para acessar esta ferramenta através do menu superior na opção *Search History*. A Tabela 14 define o cenário principal e o cenário alternativo deste caso de uso.

Tabela 14 – Caso de uso: *Select Recorded Search*

Cenário Principal:	O usuário dentro do sistema controlado do BLASTER, pode visualizar suas pesquisas préviamente salvas , podendo também visualizá-las.
Cenário Alternativo:	1. O usuário não salvou nenhuma sequência 2. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvido pelo Autor.

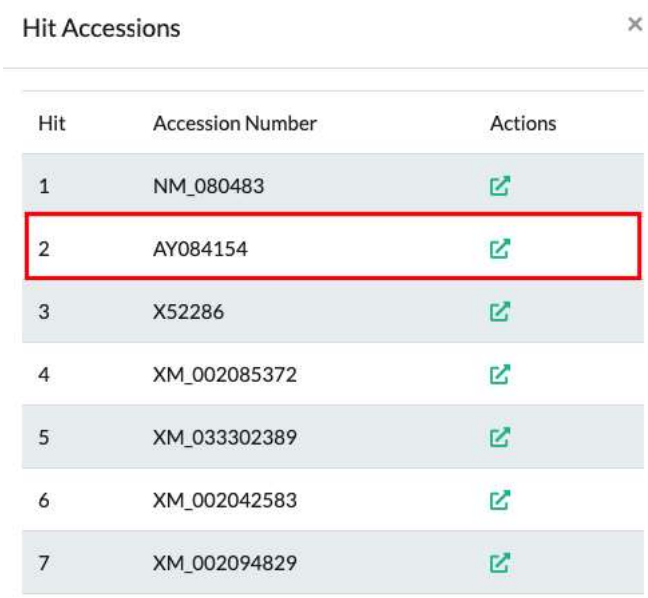
A Figura 23 exibe a tela *Search history*, onde são listadas as buscas préviamente feitas pelo pesquisador. Nesta tabela existem dados como data e hora da pesquisa, o tipo do algorimo utilizado (*type*), qual banco de dados (*database*) foi consultado, a *query* submetida e duas ações para cada registro: refazer as buscas (simbolo seta) ou visualizar (simbolo de um olho) os resultados de alinhamento relacionados a busca.

Figura 23 – *Select Recorded Search*

Fonte: Desenvolvido pelo Autor.

A ação de clicar no icone de seta nos registros da tabela da Figura 23, fará com que o BLASTER realize novamente a busca no algoritmo que foi préviamente executado e a opção *Save to my history* aparecerá desmarcada. A ação *View hit accessions* (olho), exibirá uma listagem de todos os *Hit Accessions* relacionados, como pode ser visto na Figura 24.

Figura 24 – *Hit Accessions*



Fonte: Desenvolvido pelo Autor.

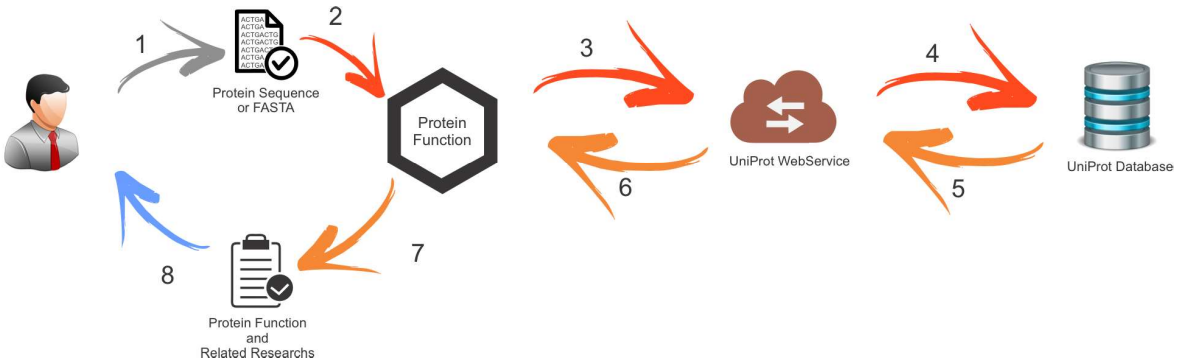
Na área demarcada da Figura 24 pode ser visto três dados: *Hit*, *Accession Number* e *Actions*. O *Hit* é o número relacionado a similadidade, quanto mais próximo de zero mais similar da *query* submetida pelo pesquisador. O *Accession Number* faz referência ao número/código do item com maior similaridade e na coluna *Actions* um *link* para o portal do NCBI, o qual exibe todos os dados do *Accession Number* selecionado.

6.2.12 *Protein Function Search*

A funcionalidade *Protein Function Search* executa a busca pela função da proteína selecionada pelo usuário. Esta busca é realizada com a passagem de um parâmetro para o servidor do Uniprot, o qual responde uma ou mais funções das proteínas listadas em sua base de dados associada ao *Accession Number* do *hit* selecionado pelo usuário ou inserido no campo *query*. Nesta funcionalidade também podem ser exibidos os textos ou artigos relacionados à *query* inserida pelo usuário.

Na Figura 25 é apresentado o fluxo completo da funcionalidade de busca da função das proteínas encontradas. Esta função pode ser acessada pelo usuário após realizado o *Search BLASTP* ou diretamente no item Protein.

Figura 25 – *Protein Function Search*



Fonte: Desenvolvido pelo Autor.

Na Tabela 15 é visualizado o cenário principal e os alternativos deste caso de uso.

Tabela 15 – Caso de uso: *Protein Function Search*

Cenário Principal:	O usuário pode visualizar a lista de funções de uma proteína conhecida ao clicar no botão *SELECT* na lista de resultados do *Search BLASTP* ou inserir os dados a serem buscados no campo *query*.
Cenário Alternativo:	1. O usuário não seleciona uma sequência; 2. O usuário não preenche o campo *query*; 3. Erro de conexão por falta de acesso à Internet. 4. Não existir artigos ou textos relacionados.

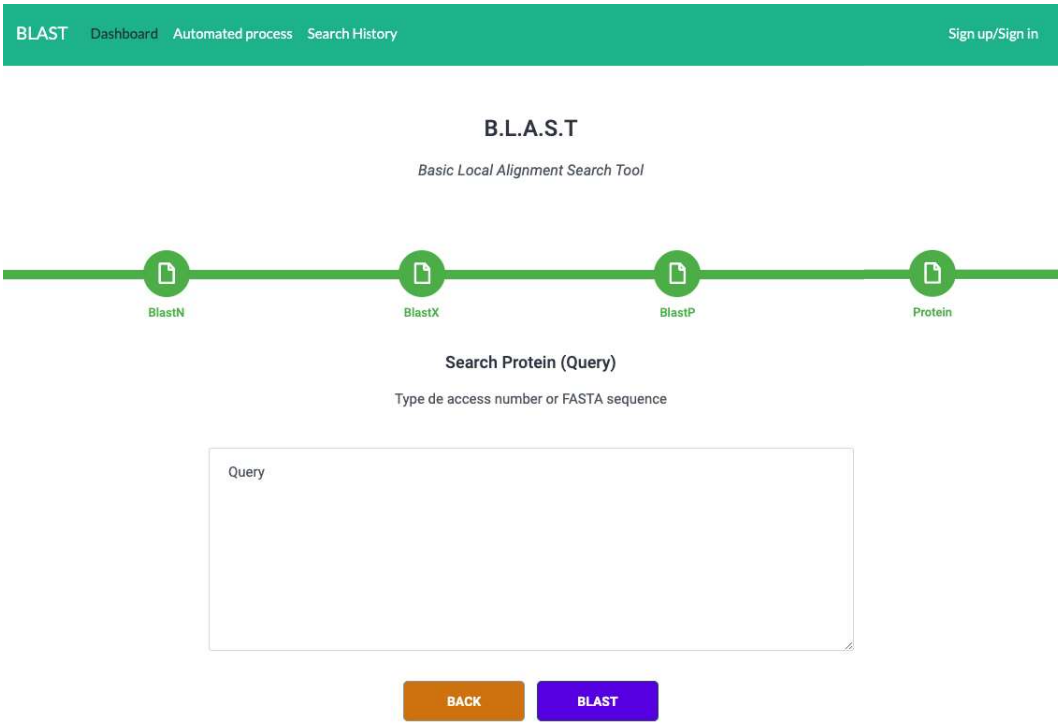
Fonte: Desenvolvido pelo Autor.

Os passos numerados 1 a 8 da Figura 25 são as funcionalidades do *Protein Function Search* descritos.

1. O usuário digita uma sequência de aminoácidos, dados no formato FASTA ou um *Accession Number*;
2. Os dados são submetidos ao *Protein Function*;
3. Envio aos *WebService* do Uniprot;
4. Consulta a base de dados do UniPtot;
5. Retorno dos dados pesquisados;
6. Envio do XML gerado ao BLASTER;
7. Interpretação e montagem da tabela dinâmica com os dados obtidos;
8. Exibição dos dados obtidos ao usuário.

A Figura 26 representa a tela na qual o usuário deverá inserir os dados a serem procurados. Esta tela também pode ser acessada através do botão *SELECT* nos resultados da busca *Search BlastP*. Os dados necessários para esta busca são: o nome do organisno, uma sequência no formato FASTA ou o *Accession Number*.

Figura 26 – *Protein Function Search*



Fonte: Desenvolvido pelo Autor.

Da mesma maneira que são executadas as ações no *Search BLASTN* de forma assíncrona, nesta funcionalidade também são executadas utilizando AJAX, a interpretação dos dados obtidos do servidor do UniProt é realizada utilizando PHP baseada na leitura do arquivo XML recebido e a montagem da tabela dinâmica é realizada com *jQuery* para ser apresentada ao usuário.

6.2.13 *View Protein Function*

View Protein Function é uma funcionalidade que exibe a função da proteína em si. Esta funcionalidade vem seguida de toda informação da proteína, gene, organismo, composição química, família, domínio e outras proteínas similares. Na Tabela 16 é apresentado o cenário principal e os alternativos deste caso de uso.

Tabela 16 – Caso de uso: *View Protein Function*

Cenário Principal:	O usuário pode visualizar a funções de uma proteína conhecida ao clicar no código na coluna *Entry*.
Cenário Alternativo:	1. Nenhuma função de proteína foi encontrada; 2. Erro de conexão por falta de acesso à Internet.

Fonte: Desenvolvido pelo Autor.

6.2.14 *View Research*

A funcionalidade *View Research* exibe os artigos ou textos selecionados na lista exibida pela *Search Recent Research*. Na Tabela 17 é apresentado o cenário principal e os alternativos deste caso de uso.

Tabela 17 – Caso de uso: *View Research*

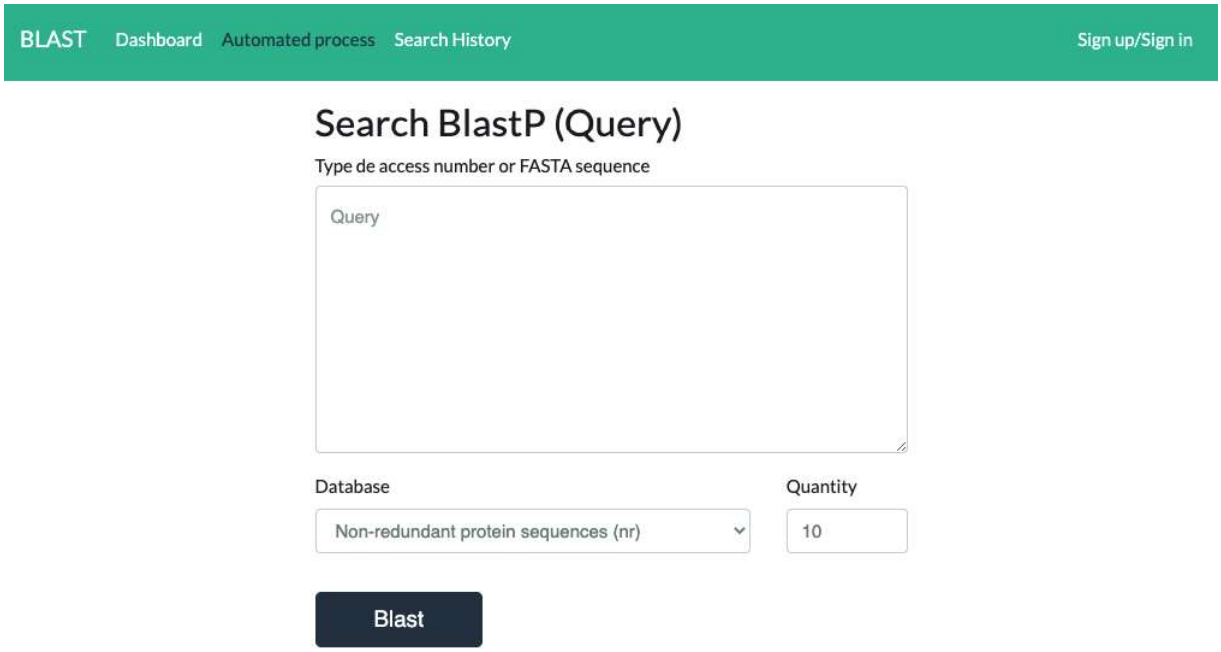
Cenário Principal:	O usuário pode visualizar os artigos e textos relacionados ao que foi pesquisado .
Cenário Alternativo:	1. Erro de conexão por falta de acesso à Internet. 2. Não existir artigos ou textos relacionados.

Fonte: Desenvolvido pelo Autor.

6.2.15 *Automated Process*

Nesta Seção será apresentado um processo automatizado do sistema completo. Este processo automatizado é a junção do *Search BLASTP* e o *Protein* e inicia-se quando o usuário submete dentro do BLASTER uma sequência de aminoácidos, um *Acession Number* ou uma sequência no formato FASTA, o processo automatizado inicia-se quando o usuário pressiona o botão BLAST. Este processo automático pode ser acessado a qualquer momento através do menu *Automated Process* (Figura 27) na barra superior, o qual será executada uma consulta nos servidores e bases de dados do portal NCBI e Uniprot. Este processo pode demorar vários minutos, pois há um grande fluxo de dados transitando em segundo plano por meio da execução de *scripts* AJAX.

Figura 27 – *Automated Process*



Fonte: Desenvolvido pelo Autor.

Ainda na Figura 27, o usuário deve selecionar a quantidade de resultados no campo *Quantity*. Este valor refere-se à quantidade de resultados que o processo automatizado resultará de possíveis *Matches*[5], os alinhamentos similares. Após estes processamentos, é realizada uma busca pela função de proteínas, onde serão consultados os servidores e bases de dados do portal UniProt. Por último, serão exibidos textos, artigos e livros relacionados aos dados obtidos. Este processo se dá por consulta ao servidor e à base de dados do portal NCBI passando como parâmetro o *Accession number*. Na Tabela 18 é apresentado o cenário principal e o alternativo deste caso de uso.

Tabela 18 – Caso de uso: *Automated Process*

Cenário Principal:	O usuário pode automatizar o processo de busca inserindo a *query* no formulário e selecionar o botão *BLAST*.
Cenário Alternativo:	1. Erro de conexão por falta de acesso à Internet; 2. O usuário não preenche o campo *query*; 3. O usuário insere uma sequência com poucos pares de base; 4. Não é encontrada nenhuma sequência similar.

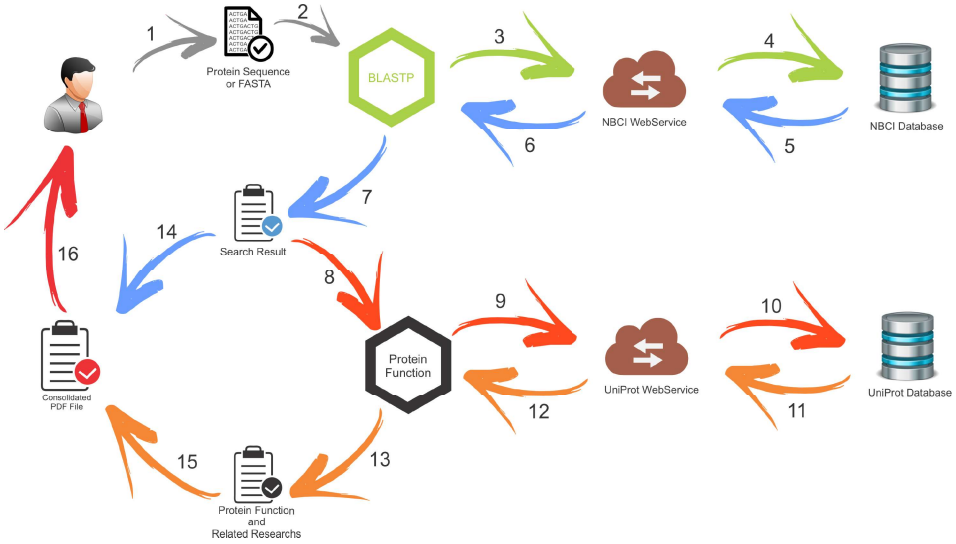
Fonte: Desenvolvida pelo Autor.

Como pode ser visto na Figura 28, existem 16 passos para a geração de um arquivo no formato PDF com todos os dados consolidados. A interação do usuário nesse processo ocorre

[5] Identidades

somente nos passos 1 e 2 já descritos na Seção 6.2.6.3, as ações de 3 a 7 e 8 a 13, estão descritas nas Seções 6.2.6.3, 6.2.12 respectivamente.

Figura 28 – *Automated Process*



Fonte: Desenvolvido pelo Autor.

Ainda na Figura 28, as ações 14 e 15 farão a inserção dos dados dentro de um único arquivo, utilizando um algoritmo de *append*. Por último, a ação 16 é a entrega do arquivo PDF com os dados consolidados ao usuário.

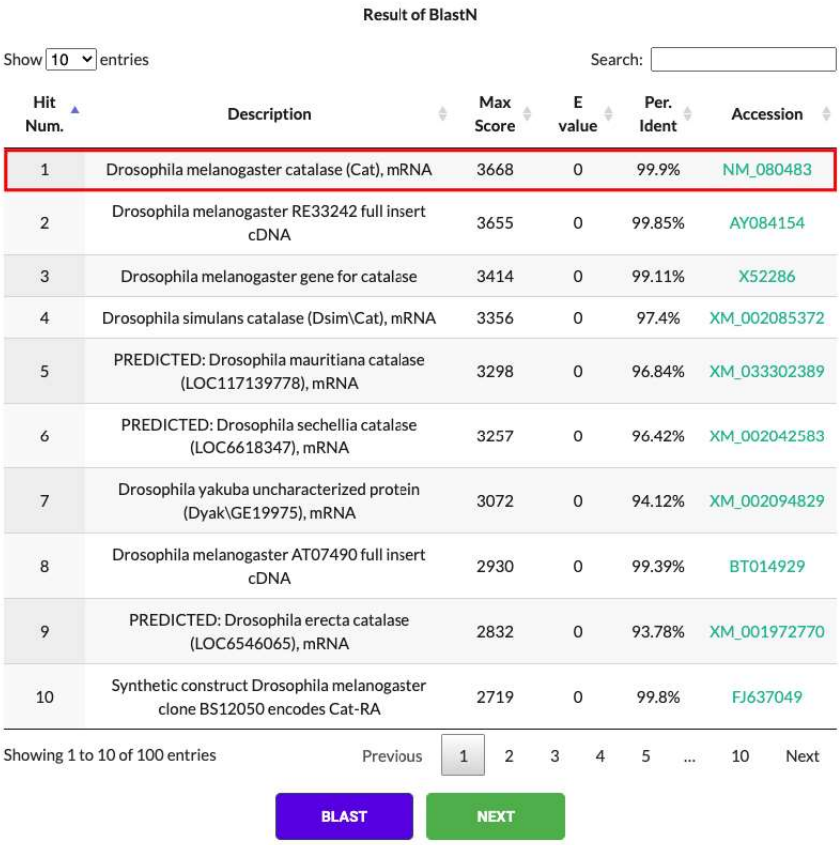
7 RESULTADOS

Nesta seção serão apresentados os resultados obtidos com este estudo, sendo eles, os resultados do *Search BlastN*, *Search BlastN*, *Search BlastP*, *Search Protein* e *Automated Process* respectivamente.

7.1 RESULTADOS *SEARCH BLASTN*

Na Figura 29, é apresentado um fragmento dos resultados obtidos pelo *Search BLASTN* descrito na Seção 6.2.6.1, sendo mostrado apenas 10 resultados de 100 resultados encontrados. Contemplando *Hit Num.* como número identificador da sequência com similaridade ao que foi consultado (este número é sequencial e o menor número corresponde a uma maior similaridade), *Description* nome da sequência, *Max Score* identifica a quantidade de nucleotídeos testados e encontrados (este dado é referente a uma nota atribuída para cada similaridade), o *E Value (Expect Value)* que significa a confiabilidade de semelhan ça, representado em porcentagem, dentro de uma busca em uma base de dados, *Per. Ident* exibe o percentual de identidade entre a sequência a ser buscada e o alvo e por último *Accession* que mostra um código que faz referência a uma página dentro do portal do NCBI onde são exibidas todas as informações do resultado selecionado. Qualquer um dos códigos *Accession* retornados do XML e apresentados na tabela, podem ser consultados no site do NCBI.

Figura 29 – Resultado da Consulta: *Search BLASTN*



Result of BlastN

Show 10 entries Search:

Hit Num.	Description	Max Score	E value	Per. Ident	Accession
1	Drosophila melanogaster catalase (Cat), mRNA	3668	0	99.9%	NM_080483
2	Drosophila melanogaster RE33242 full insert cDNA	3655	0	99.85%	AY084154
3	Drosophila melanogaster gene for catalase	3414	0	99.11%	X52286
4	Drosophila simulans catalase (Dsim\Cat), mRNA	3356	0	97.4%	XM_002085372
5	PREDICTED: Drosophila mauritiana catalase (LOC117139778), mRNA	3298	0	96.84%	XM_033302389
6	PREDICTED: Drosophila sechellia catalase (LOC6618347), mRNA	3257	0	96.42%	XM_002042583
7	Drosophila yakuba uncharacterized protein (Dyak\GE19975), mRNA	3072	0	94.12%	XM_002094829
8	Drosophila melanogaster AT07490 full insert cDNA	2930	0	99.39%	BT014929
9	PREDICTED: Drosophila erecta catalase (LOC6546065), mRNA	2832	0	93.78%	XM_001972770
10	Synthetic construct Drosophila melanogaster clone BS12050 encodes Cat-RA	2719	0	99.8%	FJ637049

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 ... 10 Next

BLAST NEXT

Fonte: Desenvolvido pelo Autor.

Ainda na Figura 29, é demonstrada uma área marcada em vermelho no *Hit Num.* 1, o usuário poderá interagir, com um clique, visualizando o alinhamento do *match* listado.

Na Figura 30 é exibido o alinhamento de uma das sequências que o usuário selecionou. Esta tela é compilada a partir da leitura do arquivo XML retornado do servidor do NCBI, interpretado com a linguagem PHP e exibida com JavaScript ao usuário. Todos os alinhamentos da *query* digitada pelo usuário e do *Sbjt (Subject*, sequência com maior similaridade encontrada na base do NCBI) são feitos com o PHP.

Figura 30 – Fragmento do Alinhamento: *View Sequence Alignment of BLASTN*

Fonte: Desenvolvido pelo Autor.

7.2 RESULTADOS *SEARCH BLASTX*

Diferente dos resultados apresentados na Seção 7.1, o *Search BLASTX* retorna um dado a mais que o *Search BLASTN*, o código FASTA da sequência alvo. Também é apresentada uma coluna denominada *Action* com um botão *SELECT* em cada um dos *Hits* encontrados, como é apresentado na Figura 31. Nestes resultados pode ser visto uma coluna FASTA, o qual exibe um *link* onde o usuário poderá ver a sequência FASTA encontrada.

Figura 31 – Fragmento do resultado da consulta: *Search BLASTX*

Result BlastX

Show 10 entries Search:

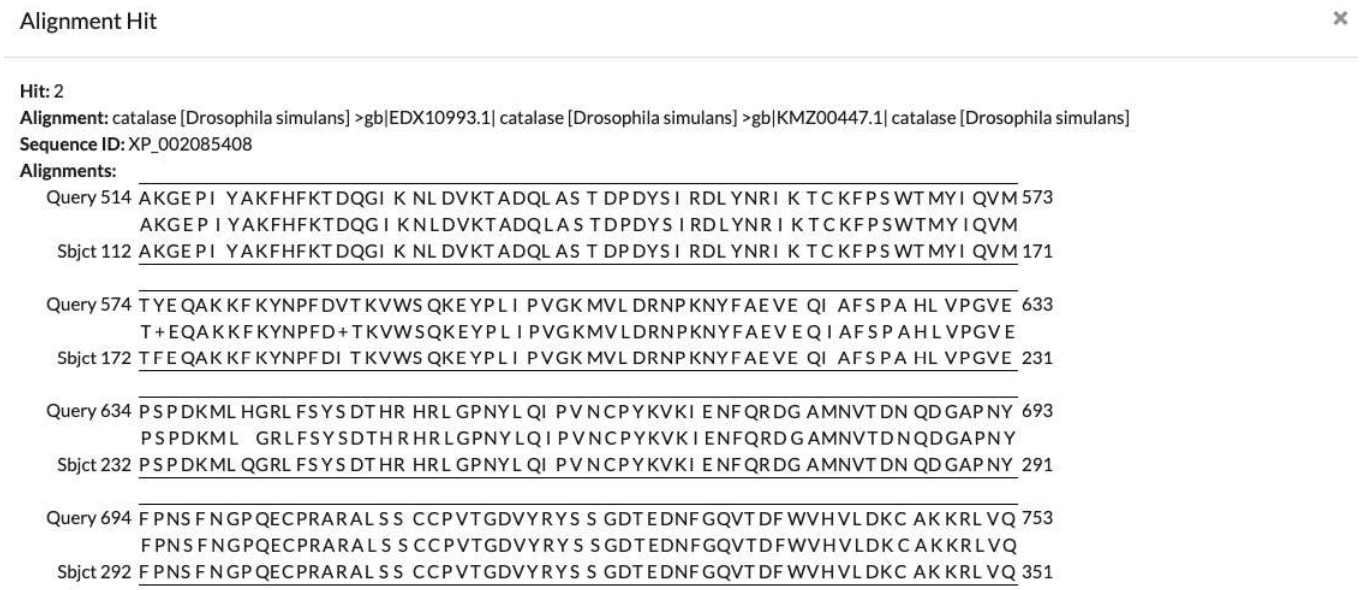
Hit Num.	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession	FASTA	Action
1	catalase [Drosophila melanogaster] >sp|P17336.2| RecName: Full=Catalase [Drosophila melanogaster] >gb|ACL91308.1| Cat-PA [synthetic construct] >gb|AAC13738.1| catalase [Drosophila melanogaster] >gb|AAF49228.1| catalase [Drosophila melanogaster] >gb|AAL89892.1| RE33242p [Drosophila melanogaster] >gb|ACL92618.1| Cat-PA, partial [synthetic construct]	833	1060	78%	0	100	NP_536731	NP_536731	SELECT
2	catalase [Drosophila simulans] >gb|EDX10993.1| catalase [Drosophila simulans] >gb|KMZ00447.1| catalase [Drosophila simulans]	825	1050	77%	0	98.98	XP_002085408	XP_002085408	SELECT
3	catalase [Drosophila mauritiana]	825	1050	77%	0	99.23	XP_033158280	XP_033158280	SELECT
4	catalase [Drosophila sechellia] >gb|EDW46546.1| Cat [Drosophila sechellia]	825	1049	77%	0	98.98	XP_002042619	XP_002042619	SELECT

Fonte: Desenvolvido pelo Autor.

Na Figura 31 o botão *SELECT* tem a função de capturar a sequência FASTA representado pela coluna FASTA e continuar para o *Search BLASTP*, preenchendo o campo *query* do próximo estado do *pypeline*.

A área demarcada na Figura 31 representa a área de clique do mouse, onde o usuário poderá interagir, para visualizar o resultado do alinhamento, exemplificado na Figura 32. Estes dados exibidos são os alinhamentos de uma das sequências que o usuário selecionou. Esta tela foi compilada com a leitura do arquivo XML recebido como resposta da consulta ao servidor do NCBI, interpretado com a linguagem PHP e exibida com JavaScript ao usuário.

Figura 32 – Fragmento do Alinhamento: *View Sequence Alignment of BLASTX*



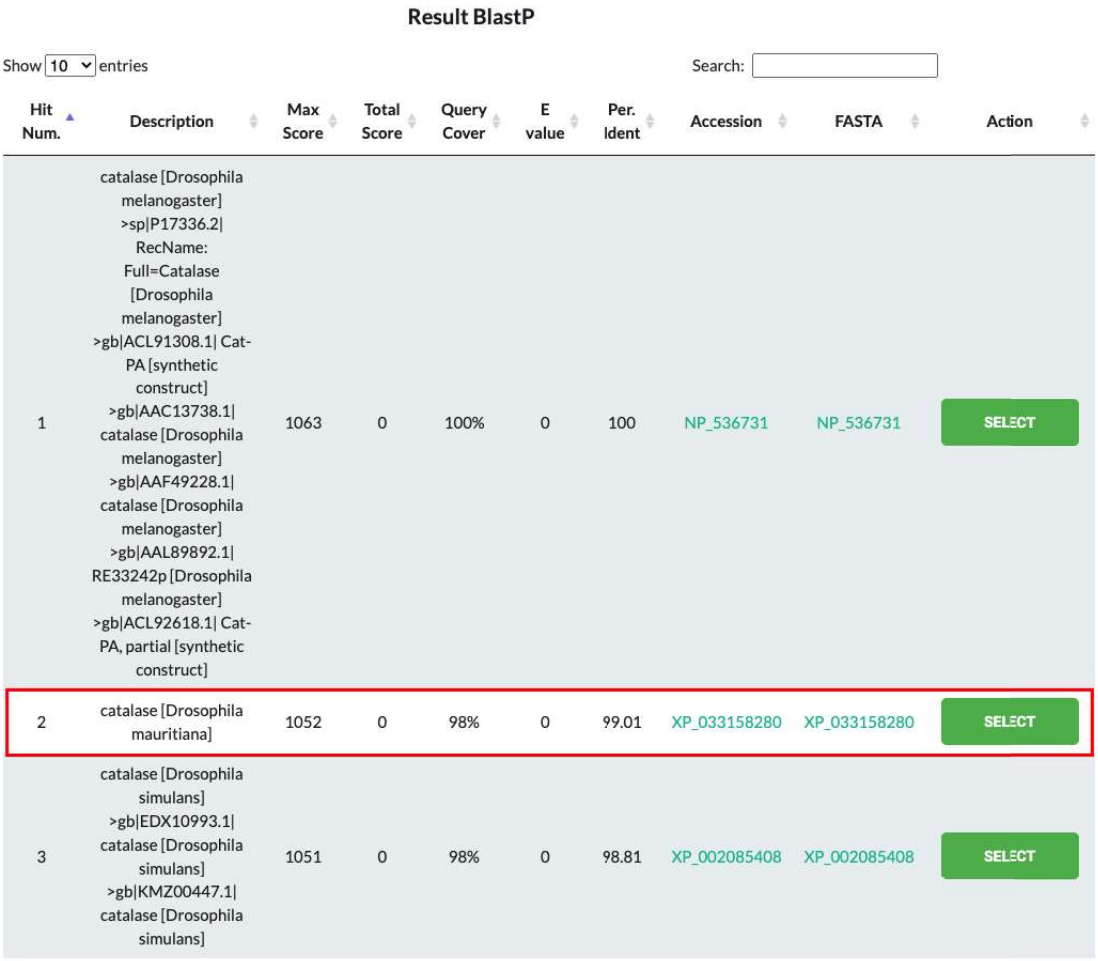
Fonte: Desenvolvido pelo Autor.

7.3 RESULTADOS *SEARCH BLASTP*

Da mesma forma, que foi apresentado nas Figuras 29 e 31 a listagem de resultados do *Search BLASTP* pode ser visualizada na Figura 33. É observado também uma área demarcada em vermelho onde pode haver interação do usuário (clique na linha do *Hit* escolhido).

Em cada um dos Hit Num. listados na Figura 33 observa-se um botão *SELECT*, o qual de levar o usuário à próxima etapa do BLASTER com os dados FASTA contidos na coluna FASTA da listagem.A funcionalidade do clique do *mouse* representado pela área demarcada na Figura 33, apresenta um resultado de alinhamento entre a *query* submetida pelo usuário e a sequência alvo.

Figura 33 – Fragmento do resultado da consulta: *Search BLASTP*



Result BlastP

Show 10 entries Search:

Hit Num.	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession	FASTA	Action
1	catalase [Drosophila melanogaster] >sp|P17336.2| RecName: Full=Catalase [Drosophila melanogaster] >gb|ACL91308.1| Cat-PA [synthetic construct] >gb|AAC13738.1| catalase [Drosophila melanogaster] >gb|AAF49228.1| catalase [Drosophila melanogaster] >gb|AAL89892.1| RE33242p [Drosophila melanogaster] >gb|ACL92618.1| Cat-PA, partial [synthetic construct]	1063	0	100%	0	100	NP_536731	NP_536731	SELECT
2	catalase [Drosophila mauritiana]	1052	0	98%	0	99.01	XP_033158280	XP_033158280	SELECT
3	catalase [Drosophila simulans] >gb|EDX10993.1| catalase [Drosophila simulans] >gb|KMZ00447.1| catalase [Drosophila simulans]	1051	0	98%	0	98.81	XP_002085408	XP_002085408	SELECT

Fonte: Desenvolvido pelo Autor.

Na Figura 34 é exibido o alinhamento de uma das sequências que o usuário pode interagir. Por sua vez, na coluna FASTA, há um *link*, que exibe ao usuário a sequência FASTA relacionada ao *Hit* selecionado.

Figura 34 – Fragmento do Alinhamento: *View Sequence Alignment of BLASTP*

Fonte: Desenvolvido pelo Autor.

Como pode ser observado na Figura 34, diferente do *View Sequence Alignment of BLASTN*, o *Search BlastP* exibe o alinhamento de uma sequência de aminoácidos.

7.4 RESULTADO *SEARCH PROTEIN* E *RELATED RESEARCH*

No processo de consulta à funcionalidade *Search Protein* obtem-se resultados exemplificados na Figura 35. Neste fragmento de um resultado de busca, pode-se observar duas áreas demarcadas: A - contendo as informações encontradas que são relacionadas a *query* submetida pelo usuário e B contendo as bibliografias relacionadas a *query* submetida.

Figura 35 – Resultado da Consulta: *Search Protein*

Fonte: Desenvolvido pelo Autor.

Ainda na Figura 35, observa-se na área demarcada A que há dois *links*: *Organism* e *NCBI Taxonomy*. *Organism* faz referência ao portal no Uniprot e *NCBI Taxonomi* faz referência ao portal do NCBI respectivamente. Ainda na área demarcada Figura 35 A observa-se a função da proteína (*Function*), atividade calalítica (textitCatalytic Activity) e o cofator (*Cofactor*) a qual a proteína representa.

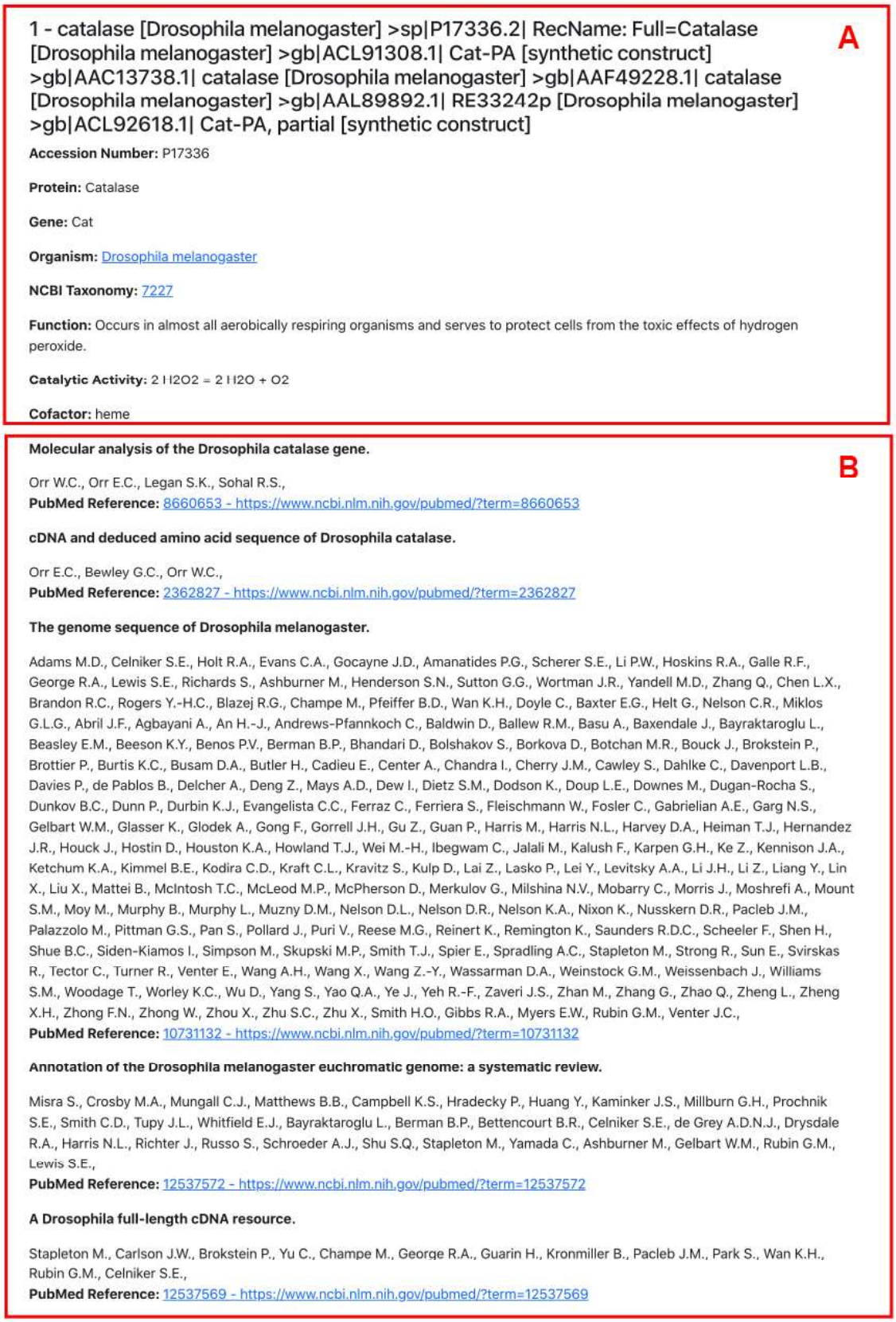
Na área demarcada (Figura 35 B) pode-se observar que os itens 1 e 2 não fazem referências diretas ao portal do PubMed[1] e os itens 3 e 4 apresentam links os quais vão ligar o pesquisador diretamente ao texto relacionado.

7.5 RESULTADO *AUTOMATED PROCESS*

Os resultados apresentados na Figura 36 são semelhantes ao apresentado na Figura 35, diferenciando somente na apresentação da descrição do *hit* encontrado.

[1] Os resultados de referências provêm do portal do Uniptot no arquivo XML recebido durante a consulta, quando não há nenhuma referência não é exebido para o usuário o *link* para acessar a bibliografia.

Figura 36 – Fragmento dos Resultado: *Automated Process*



Fonte: Desenvolvido pelo Autor.

A área demarcada A da Figura 36 apresenta as informações encontradas nos sistemas do BLASTP do NCBI e no portal do Uniprot, contendo o *Accession Number*, *Protein*, *Gene*, *NCBI Taxonomy*, *Function*, *Catalytic Activity* e *Cofactor*. Ainda na Figura 36, a área B exibe as bibliografias catalogadas relacionadas com o *hit* encontrado, todos os *links* apresentados fazem referência ao portal do NCBI (PUBMed). Nestes *links* o usuário pode clicar (quando estiver visualizando em tela) ou digitar a *Uniform Resource Locator* (URL) no navegador de internet.

Todos os dados encontrados são compilados e exibidos na tela ao usuário, tendo a necessidade, pode-se gerar um arquivo PDF, deixando opcional ao usuário salvar, imprimir ou apenas visualisar os dados obtidos. Neste processo automatizado pode não ser encontrado *Hits*, o que apresentaria para o usuário a frase *No Matches Found*, não havendo a possibilidade de geração em tela ou impressão/salvamento do arquivo PDF consolidado.

8 CONCLUSÕES

O desenvolvimento deste trabalho proporciona aos pesquisadores uma forma eficiente de obter as informações necessárias para pesquisas em um só local. A ferramenta intitulada BLASTER foi desenvolvida com objetivo de agregar duas plataformas (o BLAST e o Uniprot), facilitando a obtenção de dados pelo pesquisador. A junção dessas duas ferramentas torna possível ao pesquisador obter dados de sequências já depositadas, funções de proteínas e textos relacionados, em apenas um local. Este compilado de informações é entregue ao usuário para visualização com a possibilidade de imprimir ou salvar os dados no formato PDF.

Pelo exposto, pode-se concluir que a utilização do BLASTER, o qual agrega ferramentas distintas, diminui o trabalho por parte do usuário/pesquisador tornando assim viável sua utilização.

Como sugestão a trabalhos futuros, que podem ser desenvolvidos utilizando o BLASTER, como desenvolver um aplicativo para dispositivo móvel que pode ser acessado de qualquer lugar e a qualquer momento pelo pesquisador.

REFERÊNCIAS

ALBERTS, B. *et al. Fundamentos da biologia celular*. [S.l.]: Artmed Editora, 2002.

ALTSCHUL, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, v. 25, n. 17, p. 3389–3402, set. 1997. ISSN 13624962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/25.17.3389>.

CELESTI, F. *et al.* optimizing the research of dna sequences in a nosql document database: A preliminary study. In: IEEE. *2019 IEEE Symposium on Computers and Communications (ISCC)*. [S.l.], 2019. p. 1153–1158.

CLAVERIE, J.-M.; NOTREDAME, C. *Bioinformatics for dummies*. [S.l.]: John Wiley & Sons, 2006.

CLUSTALW/CLUSTALX. 2020. Disponível em: <http://www.clustal.org/clustal2/>. Acesso em: 31 de janeiro de 2020.

CRICK, F. Central dogma of molecular biology. *Nature*, Nature Publishing Group, v. 227, n. 5258, p. 561–563, 1970.

CRICK, F. H. On protein synthesis. In: *Symp Soc Exp Biol*. [S.l.: s.n.], 1958. v. 12, n. 138-63, p. 8.

EDGAR, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, v. 5, p. 113, ago. 2004. ISSN 1471-2105.

EL-METWALLY, S.; OUDA, O. M.; HELMY, M. *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*. [S.l.]: Springer Science & Business, 2014. Google-Books-ID: APG7BAAAQBAJ. ISBN 9781493907151.

GU, J.; BOURNE, P. E. *Structural bioinformatics*. [S.l.]: John Wiley & Sons, 2009. v. 44.

HASHEMIFAR, S. *et al.* Modulealign: module-based global alignment of protein–protein interaction networks. *Bioinformatics*, Oxford University Press, v. 32, n. 17, p. i658–i664, 2016.

HUNG, C.-L. *et al.* CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs. *Computational Biology and Chemistry*, v. 58, p. 62–68, out. 2015. ISSN 14769271. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1476927115300098>.

JARARWEH, Y. *et al.* Improving the performance of the needleman-wunsch algorithm using parallelization and vectorization techniques. *Multimedia Tools and Applications*, v. 78, n. 4, p. 3961–3977, fev. 2019. ISSN 1573-7721. Disponível em: <https://doi.org/10.1007/s11042-017-5092-0>.

JOHNSON, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Research*, v. 36, n. Web Server, p. W5–W9, maio 2008. ISSN 0305-1048, 1362-4962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn201>.

KATOH, K.; TOH, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, v. 9, n. 4, p. 286–298, mar. 2008. ISSN 1467-5463, 1477-4054. Disponível em: <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbn013>.

KUMAR, S. *et al.* MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, v. 35, n. 6, p. 1547–1549, jun. 2018. ISSN 0737-4038, 1537-1719. Disponível em: <https://academic.oup.com/mbe/article/35/6/1547/4990887>.

KUMAR, S.; TAMURA, K.; NEI, M. Mega3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in bioinformatics*, Oxford University Press, v. 5, n. 2, p. 150–163, 2004.

LESK, A. *Introduction to bioinformatics*. [S.l.]: Oxford university press, 2019.

MEGA10. 2020. Disponível em: <https://www.megasoftware.net/docs>. Acesso em: 01 de fevereiro de 2020.

MOUNT, D. W. Using PAM Matrices in Sequence Alignments. *Cold Spring Harbor Protocols*, v. 2008, n. 6, p. pdb.top38–pdb.top38, jun. 2008. ISSN 1559-6095. Disponível em: <http://www.cshprotocols.org/cgi/doi/10.1101/pdb.top38>.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, v. 48, n. 3, p. 443–453, mar. 1970. ISSN 00222836. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0022283670900574>.

NELSON, D. L.; COX, M. M. *Princípios de Bioquímica de Lehninger-7*. [S.l.]: Artmed Editora, 2018.

NORMAND, A. *et al.* Nucleotide sequence database comparison for internal transcribed spacer 2 genetic region dna barcode dermatophyte routine identification. *Journal of Clinical Microbiology*, Am Soc Microbiol, 2018.

NOTREDAME, C. Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLOS Computational Biology*, v. 3, n. 8, p. e123, ago. 2007. ISSN 1553-7358. Disponível em: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030123>.

NOTREDAME, C.; HIGGINS, D. G.; HERINGA, J. T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1edited by J. Thornton. *Journal of Molecular Biology*, v. 302, n. 1, p. 205–217, set. 2000. ISSN 00222836. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0022283600940427>.

OLIVER, T. *et al.* Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics*, v. 21, n. 16, p. 3431–3432, ago. 2005. ISSN 1367-4803, 1460-2059. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti508>.

PAIS, F. S.-M. *et al.* Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*, v. 9, n. 1, p. 4, 2014. ISSN 1748-7188. Disponível em: <http://almob.biomedcentral.com/articles/10.1186/1748-7188-9-4>.

ROZEWICKI, J. *et al.* MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research*, p. gkz342, maio 2019. ISSN 0305-1048, 1362-4962. Disponível em: <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz342/5486273>.

SAUVAGE, T. *et al.* Tree2fasta: a flexible perl script for batch extraction of fasta sequences from exploratory phylogenetic trees. *BMC research notes*, BioMed Central, v. 11, n. 1, p. 164, 2018.

SETUBAL, J. C.; MEIDANIS, J.; SETUBAL-MEIDANIS, . *Introduction to computational molecular biology*. [S.l.]: PWS Pub. Boston, 1997.

SIEVERS, F. *et al.* Making automated multiple alignments of very large numbers of protein sequences. *Bioinformatics*, v. 29, n. 8, p. 989–995, abr. 2013. ISSN 1460-2059, 1367-4803. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt093>.

SIEVERS, F.; HIGGINS, D. G. Clustal Omega for making accurate alignments of many protein sequences: Clustal Omega for Many Protein Sequences. *Protein Science*, v. 27, n. 1, p. 135–145, jan. 2018. ISSN 09618368. Disponível em: <http://doi.wiley.com/10.1002/pro.3290>.

SMITH, T. F.; WATERMAN, M. S. *et al.* Identification of common molecular subsequences. *Journal of molecular biology*, Elsevier Science, v. 147, n. 1, p. 195–197, 1981.

SOLOMON, J. *et al.* Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, ACM New York, NY, USA, v. 35, n. 4, p. 1–13, 2016.

SZKLARCZYK, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. v. 39, p. D561–D568, 2011. ISSN 0305-1048, 1362-4962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq973>.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, v. 22, n. 22, p. 4673–4680, nov. 1994. ISSN 0305-1048.

TOMMASO, P. D. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research*, v. 39, n. suppl, p. W13–W17, jul. 2011. ISSN 0305-1048, 1362-4962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr245>.

WATSON, J. D. *DNA recombinante: genes e genomas*. Porto Alegre (RS): ARTMED, 2009. OCLC: 817218630. ISBN 978-85-363-1375-7.

WATSON, J. D. *et al. Biologia molecular do gene*. [S.l.]: Artmed Editora, 2015.