

Pedro Azevedo Abrantes de Oliveira
Orientador: Romuere Rodrigues Veloso e Silva

GeMapCom 2.0: Uma Ferramenta para Mapeamento Comparativo e Visualização Interativa de Dados Genômicas

Picos - PI
19 de fevereiro de 2023

Pedro Azevedo Abrantes de Oliveira
Orientador: Romuere Rodrigues Veloso e Silva

GeMapCom 2.0: Uma Ferramenta para Mapeamento Comparativo e Visualização Interativa de Dados Genômicas

Trabalho de Conclusão de Curso de Bacharelado em Sistemas de Informação do Campus Senador Helvídio Nunes de Barros da Universidade Federal do Piauí como parte dos requisitos para obtenção do Grau de Bacharel em Sistemas de Informação.

Universidade Federal do Piauí
Campus Senador Helvídio Nunes de Barros
Bacharelado em Sistemas de Informação

Picos - PI
19 de fevereiro de 2023

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca José Albano de Macêdo

O48g Oliveira, Pedro Azevedo Abrantes de
GeMapCom 2.0 : uma ferramenta para mapeamento comparativo e
visualização interativa de dados genômicas [recurso eletrônico] / Pedro
Azevedo Abrantes de Oliveira – 2023.
41 f.

1 Arquivo em PDF

Indexado no catálogo *online* da biblioteca José Albano de Macêdo-CSHNB
Aberto a pesquisadores, com restrições da Biblioteca

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do
Piauí, Bacharel em Sistemas de Informação, Picos, 2023.

“Orientador: Dr. Romuere Rodrigues Veloso e Silva”

1. Alinhamento genético. 2. Bioinformática. 3. Genômica Comparativa. 4.
Visualização Genômica. I. Silva, Romuere Rodrigues Veloso e. II. Título.

CDD 572.8

GEMAPCOM 2.0: UMA FERRAMENTA PARA MAPEAMENTO COMPARATIVO E
VISUALIZAÇÃO INTERATIVA DE DADOS GENÔMICAS

PEDRO AZEVEDO ABRANTES DE OLIVEIRA

Monografia aprovada como exigência parcial para obtenção do grau de Bacharel em Sistemas
de Informação.

Data de Aprovação

Picos – PI, 03 de março de 2023



Prof(a). Dr. Romuere Rodrigues Veloso e Silva



Prof(a). Dr. Flávio Henrique Duarte de Araújo



Prof(a). Me. Júlio Vitor Monteiro Marques

Agradecimentos

Quero agradecer primeiramente a Deus, por estar me concedendo a oportunidade de concluir o curso que sempre sonhei em fazer e por conceder a realização deste trabalho. A Ele, toda a gratidão.

Aos meus pais, Robson Abrantes de Oliveira e Maria Dalva Azevedo Abrantes de Oliveira por ser a base de tudo, pelo incentivo e trabalho de ambos cheguei até aqui. Aos meus irmãos Rafaela Azevedo Abrantes de Oliveira, Andreia Azevedo Abrantes de Oliveira e Arthur Azevedo Abrantes de Oliveira que sempre me ajudaram nos momentos mais difíceis que passei nessa caminhada.

Agradecer à Universidade Federal do Piauí por fornecer a estrutura necessária durante o período de graduação.

Agradecer meu orientador, Prof. Dr. Romuere Rodrigues Veloso e Silva, por me guiar desde o início da minha pesquisa. Obrigado pela positividade, paciência e ensinamentos transmitidos e pelo grande exemplo de determinação oferecendo oportunidades de crescimento acadêmico.

Aos meus professores da graduação, muito obrigado. Aos amigos Narciso de Sousa Rodrigues e Saul Sousa da Rocha, companheiros de trabalhos e irmãos na amizade que fizeram parte da minha formação, obrigado.

Por fim, gostaria de expressar minha gratidão a Marta Maria Cordeiro, uma amiga incrível e minha amada namorada, obrigado por esta sempre comigo me ajudando independente da situação. Obrigado por ser uma das peças chaves da minha formação!

Somos o que pensamos. Tudo o que somos surge com nossos pensamentos. Com nossos pensamentos, fazemos o nosso mundo.

Buda

Resumo

A bioinformática é uma área interdisciplinar que tem como objetivo aplicar tecnologias computacionais para compreender os sistemas biológicos. Este campo tem crescido enormemente com o avanço na tecnologia de sequenciamento de DNA, que permitiu a obtenção de grandes quantidades de informações genômicas. Com o aumento no número de genomas sequenciados, houve também uma necessidade de desenvolver bancos de dados especializados para armazenar essas informações e ferramentas para comparar e visualizar esses dados. Muitas das ferramentas existentes apresentam limitações, como informações incompletas na comparação de genes entre diferentes espécies e poucos recursos de interação com usuários. É neste contexto que o trabalho de desenvolvimento da ferramenta GeMapCom se insere. A GeMapCom é uma ferramenta para o mapeamento comparativo e visualização de dados genômicos, que tem como objetivo fornecer um produto totalmente interativo e com diversas funcionalidades para o usuário como uma solução mais completa e intuitiva para a comparação e visualização de dados genômicos, além de facilitar a compreensão e interpretação dos dados pelo usuário.

Palavras-chaves: Alinhamento Genético; Bioinformática; Genômica Comparativa; Visualização Genômica; Genomas; BLAST; NCBI.

Abstract

Bioinformatics is a multidisciplinary field of study that combines concepts from computer science, biology, and chemistry. It arose with the aim of using computational technologies to analyze data related to molecular biology, genetics, and biochemistry. As a result of the explosion of genome studies, there has been an exponential increase in the number of sequenced and published genomes, requiring the need for specific databases to store this information and tools to compare and visualize this data. Although there are many genome comparison and visualization tools, many have incomplete information in comparing genes between different species and have few user interaction resources. This work consists of the development of the GeMapCom tool, a comparative mapping and visualization tool for genomic data, which aims to provide a tool with multiple functionalities for the user that is completely interactive. The aim is to provide a more comprehensive and intuitive solution for the comparison and visualization of genomic data, as well as to facilitate the user's understanding and interpretation of the data.

Lista de ilustrações

Figura 1 – Toda a sequência A é alinhada com toda a sequência B. Fonte: Autor. .	17
Figura 2 – Um trecho da sequência A é alinhado com um trecho da sequência B, tal que esse alinhamento é “bom”, quando comparado com outros possíveis alinhamentos locais entre A e B. Fonte: Autor.	18
Figura 3 – Ilustra uma linha de comando para a execução do BLAST+. Fonte: Autor.	18
Figura 4 – Mostra os MISMATCHES, em amarelo, os MATCHES, em verde, e os GAPS em vermelho após a comparação de duas sequências. Fonte: Autor.	19
Figura 5 – Usando Kahlammo para investigar a estrutura do gene em uma montagem do genoma de <i>Haemonchus contortus</i> . Fonte: (WINTERSINGER; WASMUTH, 2014)	21
Figura 6 – Diagrama de Caso de Uso com o Ator Interagindo com a Ferramenta GeMapCom. Fonte: Autor.	26
Figura 7 – Arquitetura da Ferramenta GeMapCom. Fonte: Autor.	27
Figura 8 – Tela Principal da Ferramenta GeMapCom. Fonte: Autor.	28
Figura 9 – Tela de Comparação da GeMapCom. Fonte: Autor.	29
Figura 10 – Tela de Conversão da GeMapCom. Fonte: Autor.	30
Figura 11 – Tela principal do Blast. Fonte: Autor.	31
Figura 12 – Tela de seleção de tipo de alinhamento do Blast. Fonte: Autor.	31
Figura 13 – Tela de busca de sequências online da ferramenta GeMapCom. Fonte: Autor.	31
Figura 14 – Tela de Filtragem do Blast. Fonte: Autor.	32
Figura 15 – Tela de Resultados da GeMapCom. Fonte: Autor.	32
Figura 16 – Tela de Resultados Gráficos do Blast. Fonte: Autor.	33
Figura 17 – Tela de Resultados Gráficos do Blast dando ênfase a possível interação do usuário com os gráficos. Fonte: Autor.	34

Lista de tabelas

Tabela 1 – Variações do Blast+ e suas descrições.	20
Tabela 2 – Trabalhos Relacionados.	23

Lista de abreviaturas e siglas

BLAST	Basic Local Alignment Search Tool
INPI	Instituto Nacional de Propriedade Industrial
NCBI	National Center for Biotechnology Information
NHGRI	National Human Genome Research Institute
RAG	Razão entre o Tamanho do Alinhamento e o Tamanho do Gene
SGBD	Sistema de Gerenciamento de Banco de Dados
UML	Unified Modeling Language

Sumário

1	Introdução	12
1.1	Objetivos Geral	13
1.2	Objetivos Específicos	13
1.3	Publicações Obtidas	13
1.4	Organização do Trabalho	14
2	Referencial Teórico	15
2.1	Genômica Comparativa	15
2.2	Banco de Dados Genômicos	16
2.3	Alinhamento de Sequência	16
2.4	BLAST+	18
2.5	Filtros	19
2.6	Kablammo	21
3	Trabalhos Relacionados	22
4	Metodologia Proposta	25
4.1	Diagrama de caso de uso	25
4.2	Ferramentas e Tecnologias	25
4.3	Lógica de Desenvolvimento	27
5	Resultados e Discussões	28
5.1	Tela Inicial	28
5.2	Tela de Comparar Sequências	28
5.3	Tela de Conversão	29
5.4	Tela do Blast+	30
5.5	Tela de Filtros	30
5.6	Tela de Resultados	31
6	Conclusão	35
	Referências	36
	APÊNDICE A Registro de Software	38
	APÊNDICE B Certificado	40

1 Introdução

A bioinformática, conforme o [National Human Genome Research Institute \(2021\)](#), é uma subdisciplina da biologia e da informática que se encarrega da aquisição, armazenamento, análise e difusão de dados biológicos, principalmente sequências de DNA e aminoácidos. Seu objetivo principal é utilizar tecnologias computacionais para armazenar, analisar e compartilhar informações sobre genomas, proteínas e outros componentes moleculares da biologia.

O desenvolvimento da bioinformática veio a partir do aumento da quantidade de informações geradas pela sequenciação de genomas e pela investigação de processos biológicos. O acesso a tecnologias de processamento de dados mais avançadas permitiu a sequenciação de proteínas de forma automática, possibilitando aos cientistas estudar como as células se comportam e mudam durante as doenças.

No início, a bioinformática era usada apenas para estudar teoricamente os processos biológicos. Com o tempo, a evolução da tecnologia permitiu uma análise mais eficiente das sequências de proteínas. Segundo [Junior \(2002\)](#), a quantidade crescente de genomas sequenciados e publicados resultou na necessidade de análise dessas sequências para uma caracterização mais precisa das funções dos organismos estudados e seus aspectos evolutivos. Isso culminou em 2004 com a sequência completa do genoma humano. Dessa forma, a bioinformática é uma disciplina que busca melhorar a compreensão da biologia molecular e das doenças relacionadas aos processos biológicos, principalmente no que diz respeito às sequências genéticas, usando informática avançada.

Sendo assim, o sequenciamento do genoma, uma área essencial da bioinformática que estuda às sequências genéticas, permitiu um entendimento mais profundo da biologia humana e o planejamento racional de pesquisas biomédicas. Ao possibilitar a análise de todo o material genético de um organismo, o sequenciamento entusiasmou a busca por novas abordagens na determinação de riscos associados a doenças, bem como no desenvolvimento de terapias individualizadas.

Com os sequenciamentos gerados, a partir de diversas formas, começaram a existir bancos de dados universais para o armazenamento de todos esses sequenciamentos, sendo eles, o The National Center for Biotechnology Information (NCBI) ([KIM et al., 2022](#)) dos EUA, considerado o banco de dados central de informações genômicas. Integrado a ele tem-se o GenBank Sequence Database ([BENSON et al., 2015](#)) como principal banco de dados responsável por armazenar todas as sequências disponíveis publicamente de DNA, proteínas e RNA, entre outros.

Então, junto com esses bancos de dados é possível utilizar suas informações para realizar as comparações genômicas para identificar relações evolutivas entre diferentes espécies, bem como detectar a presença de variações genéticas associadas a doenças e

condições médicas. Isso é alcançado comparando a sequência genética de um organismo como referência com a sequência genética de outro organismo, localizando as semelhanças e diferenças nas sequências de DNA.

Atualmente, existem softwares limitados que realizam essas comparações que exigem muito do usuário, como uma compreensão alta de programação, ou então a necessidade de realizar treinamentos para utilizar essas ferramentas. Com o objetivo de facilitar essas limitações, surgiu a motivação para o desenvolvimento de uma ferramenta para comparação genômica dinâmico e de fácil manuseio que auxilia os biólogos a realizarem as comparações e assim obter resultados desejados.

1.1 Objetivos Geral

Este trabalho tem como objetivo principal desenvolver uma ferramenta desktop de fácil utilização, que seja interativa e apresente informações claras e organizadas. Para alcançar este objetivo, será utilizada a linguagem de programação *Python*, juntamente com a biblioteca *PySide6* e o *Basic Local Alignment Search Tool* (BLAST). O desenvolvimento dessa ferramenta possibilitará ao usuário explorar diversas funcionalidades relacionadas ao sequenciamento de genomas.

1.2 Objetivos Específicos

1. Identificar as principais ferramentas utilizadas para visualização genômica.
2. Desenvolver uma ferramenta de visualização genômica de fácil acesso.
3. Estabelecer um comparativo entre a ferramenta desenvolvida e as ferramentas já existentes acerca da visualização genômica.
4. Propiciar uma visualização interativa dos genes e dos mapeamentos comparativos.

1.3 Publicações Obtidas

OLIVEIRA, P. A. A.; SILVA, R. R. V. E. FILHO A. O. C.; ARAUJO F. H. D.; JUNIOR F. C. A. C.; SILVA L. R. G.; SARMENTO J. L. R. . GeMapCom: Uma ferramenta para mapeamento comparativo e visualização de dados Genômicos.. 2021. Patente: Programa de Computador. Número do registro: BR512021001837-0, data de registro: 01/03/2021, título: "GeMapCom: Uma ferramenta para mapeamento comparativo e visualização de dados Genômicos.", Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial. (Registro de Software ilustrado no Apêndice A).

OLIVEIRA, P. A. A.; SILVA, R. R. V. E. obtiveram o **1.º Lugar** em **TECNOLOGIAS DE PRODUÇÃO** na apresentação oral do trabalho intitulado **UMA FERRAMENTA PARA MAPEAMENTO COMPARATIVO E VISUALIZAÇÃO DE DADOS GENÔMICOS**, durante o **XIII SDTI: XIII Seminário de Desenvolvimento Tecnológico e Inovação - 2021**, realizado na UFPI – Campus Ministro Petrônio Portella – Teresina – PI. (Certificado da premiação ilustrado no Apêndice B).

1.4 Organização do Trabalho

Este trabalho é dividido em seis capítulos, organizados da seguinte maneira: Capítulo 1, apresenta a introdução. No Capítulo 2 está apresentado o referencial teórico, onde são detalhados conceitos relevantes para este trabalho. No Capítulo 3, são exibidos os principais trabalhos relacionados com esta pesquisa tecnológica. No capítulo 4, é apresentado em detalhes o desenvolvimento do GeMapCom. O Capítulo 5 apresenta o resultado final da ferramenta proposta nesta dissertação. Por último, no Capítulo 6 foram apresentadas as principais conclusões obtidas e trabalhos futuros.

2 Referencial Teórico

Neste capítulo destaca-se a relevância de se conhecer conceitos prévios que darão suporte e compreensão ao tema abordado. Desta forma, é possível obter um entendimento claro sobre a metodologia e resultados apresentados ao longo deste trabalho.

2.1 Genômica Comparativa

A genômica comparativa é uma área de pesquisa biológica que se concentra na comparação de sequências genômicas completas entre diferentes espécies ([National Human Genome Research Institute, 2020](#)). Este campo de estudo permite que os pesquisadores avaliam cuidadosamente as características que definem vários organismos e, com isso, encontrem regiões de semelhanças e diferenças ([HARDISON, 2003](#)).

O seu objetivo principal é investigar as relações entre diferentes organismos vivos, identificando similitudes e descobrindo novas informações sobre a estrutura e função dos genes, que pode ser útil para o tratamento de doenças e a obtenção de informações sobre os seres vivos ([BACHHAWAT, 2006](#)). Essa técnica é amplamente utilizada para avaliar como a localização, categorização e função dos genes variam entre diferentes espécies.

A análise comparativa permite que os cientistas detectem áreas de semelhanças e divergências entre as sequências de diferentes espécies. Isso fornece uma compreensão mais profunda da estrutura e função dos genes comparados.

Um dos impactos da genômica comparativa é a identificação de áreas de semelhanças e divergências entre os genomas de várias espécies. Por exemplo, estudos sobre a análise comparativa de diferentes tipos de animais, como vacas, cães e gatos, permitem a identificação de genes responsáveis por características específicas, como produção de leite com alto teor de gordura, problemas de sono, entre outras ([National Human Genome Research Institute, 2020](#)).

Além de aplicações na medicina molecular, a genômica comparativa também tem um impacto importante na agricultura. Estudos comparativos de genomas de plantas permitem a identificação de genes responsáveis por características agrônomicas importantes, como resistência a doenças, adaptação a condições climáticas adversas, entre outras ([SIVASHANKARI; SHANMUGHAVAL, 2007](#)).

Sendo assim, uma ferramenta valiosa para a ciência, fornecendo uma compreensão mais profunda das relações entre diferentes espécies, bem como uma fonte importante de informações para a medicina e a agricultura. Esta técnica tem a capacidade de trazer avanços significativos em áreas como cura de doenças, melhoria da produção de alimentos e compreensão da evolução molecular.

2.2 Banco de Dados Genômicos

Os bancos de dados genômicos são uma importante ferramenta para o estudo da bioinformática. Com o avanço da tecnologia, aumentou-se a produção de sequências de DNA e proteínas, o que criou a necessidade de armazenar essas informações de forma eficiente.

Antes da existência dos bancos de dados genômicos, as sequências eram armazenadas em arquivos de texto e sua manipulação era difícil. Os biólogos moleculares, então, começaram a usar Sistemas Gerenciadores de Bancos de Dados (SGBD) para gerenciar grandes volumes de dados.

Com o aumento da produção de sequências, esses dados passaram a ser submetidos aos bancos de dados genômicos através da Internet (DOOLITTLE, 1990), tornando-os acessíveis a pesquisadores em todo o mundo. Atualmente, existem vários bancos de dados genômicos importantes disponíveis, incluindo o GenBank Sequence Database (NCBI, 2018), The Integrated Microbial Genomes (JGI, 2018), EMBL Nucleotide Sequence Database (ENA, 2020), e PIR - International Protein Sequence Database (PIR, 2020).

Estes bancos de dados permitem aos pesquisadores acessar e utilizar informações genômicas de várias espécies, contribuindo para o avanço da biotecnologia e da medicina. Além disso, eles permitem a comparação de informações genômicas entre espécies, o que é fundamental para o estudo da evolução e diversidade biológica.

2.3 Alinhamento de Sequência

O alinhamento de sequências é uma técnica amplamente utilizada na biologia molecular que visa comparar duas ou mais sequências genéticas para identificar sua similaridade. O objetivo é colocar uma sequência sobre a outra de forma que a correspondência entre elas se torne evidente. O alinhamento pode ser realizado com sequências de DNA ou proteínas.

Na comparação das sequências, a similaridade é a medida do quanto as sequências são parecidas, ou seja, refere-se à porcentagem de nucleotídeos idênticos ou de aminoácidos com propriedades químicas semelhantes. A homologia, por sua vez, refere-se à relação evolutiva entre sequências. Duas sequências homólogas são derivadas de uma sequência ancestral comum.

A tarefa de alinhamento é utilizada tanto na comparação de sequências de DNA quanto na de proteínas. Para realizar o alinhamento, considera-se que as sequências são cadeias de caracteres pertencentes a determinado alfabeto. Para o DNA há o alfabeto DNA = (A, C, T, G) e para proteínas há o alfabeto PROT = (D, E, A, R, N, C, F, G, Q, H, I, L, K, M, P, S, Y, T, W, V).

As sequências que são submetidas ao alinhamento possuem tamanhos variados. Inicial-

mente coloca-se gaps, que são espaços em pontos arbitrários inseridos nas extremidades ou no interior das sequências, de modo que as sequências modificadas apresentam o mesmo comprimento. O gap é representado pelo caractere "-".

A importância dos gaps reside no fato de permitirem que cada caractere ou gap em uma sequência seja comparado a um caractere ou gap em todas as outras sequências. No modelo do alinhamento, cada par caracter-caracter ou caracter-gap recebe um peso, que será utilizado para medir a similaridade entre as sequências.

Entretanto, a inserção de gaps não contribui apenas para fazer com que as sequências resultantes possuam o mesmo tamanho. Um gap indica a ocorrência de uma possível mutação, que é a ocorrência de uma operação de inserção ou supressão de monômeros (ISAEV et al., 2004). Inserção e supressão de monômeros são conhecidos por afetar a precisão do alinhamento múltiplo, onde cada método de alinhamento escolhe a sua forma de tratar a presença de gaps (GOLUBCHIK et al., 2007).

Na genômica comparativa, há dois tipos de alinhamentos principais: global e local. O alinhamento global compara todas as sequências, resultando em uma correspondência contínua entre elas, como ilustra a Figura 1. Um exemplo desse tipo de alinhamento é o algoritmo Needleman-Wunsch, proposto por Needleman e Wunsch em 1970 (NEEDLEMAN; WUNSCH, 1970). Este algoritmo usa uma matriz de pontuação (scores) para medir a similaridade entre caracteres e procura a solução ótima. A quantidade de linhas e colunas na matriz é determinada pelo tamanho das sequências de entrada (WANG, 2002).



Figura 1 – Toda a sequência A é alinhada com toda a sequência B. Fonte: Autor.

Já no alinhamento local, as sequências são comparadas apenas nas regiões com alta similaridade, resultando em fragmentos de correspondência separados, como é mostrado na Figura 2. Um exemplo desse tipo de alinhamento é o algoritmo Smith-Waterman, desenvolvido para encontrar o alinhamento local ideal entre duas sequências (SMITH; WATERMAN et al., 1981; GOTOH, 1982). Este algoritmo é baseado em cálculos de matriz de pontuação e o número de suas linhas e colunas é determinado pelo número de resíduos nas sequências de consulta e banco de dados.

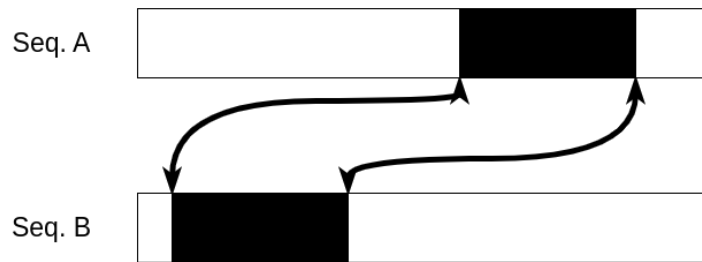


Figura 2 – Um trecho da seqüência A é alinhado com um trecho da seqüência B, tal que esse alinhamento é “bom”, quando comparado com outros possíveis alinhamentos locais entre A e B. Fonte: Autor.

2.4 BLAST+

BLAST+ é uma ferramenta de comparação de seqüências de DNA ou proteínas, desenvolvida pela *National Center for Biotechnology Information* (NCBI). É uma evolução da ferramenta original BLAST, que foi um dos primeiros programas de busca de seqüências (ALTSCHUL et al., 1990). A versão mais recente, BLAST+, oferece melhorias significativas em termos de desempenho e precisão.

O algoritmo prioriza alinhamentos de locais específicos da seqüência, com respaldo estatístico, em lugar de realizar alinhamentos “globais”, como os realizados por programas de alinhamento de múltiplas seqüências, como o Clustal2 (CHENNA et al., 2003).

O BLAST+ é uma ferramenta de linha de comando, ou seja, não tem interface gráfica de usuário como ilustra a Figura 3. Isso pode torná-lo **menos amigável para usuários menos experientes**, mas oferece mais flexibilidade e controle para usuários avançados. Além disso, a versão em linha de comando permite a execução em larga escala, tornando ideal para tarefas de análise de dados em grandes volumes.

```
pedro@pedro-Azevedo-A515-41G: ~
Arquivo Editar Ver Pesquisar Terminal Ajuda
pedro@pedro-Azevedo-A515-41G:~$ blastn -query /home/pedro/Documentos/Projeto/BLAST/Teste01.fasta -subject /home/pedro/Documentos/Projeto/BLAST/Teste02.fasta -outfmt 6 -penalty -3 -reward 2
```

Figura 3 – Ilustra uma linha de comando para a execução do BLAST+. Fonte: Autor.

A ferramenta suporta diversos tipos de bancos de dados de referência, incluindo proteínas e genomas, e permite a busca por seqüências semelhantes em múltiplos bancos de dados. Além disso, permite a definição de vários parâmetros de busca, incluindo tolerância a mutações, tipo de busca (local, global, semiautônoma) e escore.

A saída do processamento realizado pelo algoritmo BLAST consiste na criação de um alinhamento, em que as duas sequências são comparadas lado a lado. Quando os nucleotídeos ou as proteínas são iguais, é registrado um "MATCH". Quando são diferentes, é registrado um "MISMATCH", e cada MATCH e MISMATCH resulta em um aumento ou diminuição na pontuação do score final. Como mostrado na Figura 4, a comparação realizada pelo algoritmo BLAST permite visualizar a influência de cada pontuação na modificação do score final e, conseqüentemente, na escolha do melhor alinhamento.

```

Score = 2815 bits (1524), Expect = 0.0
Identities = 2498/2961 (84%), Gaps = 96/2961 (3%)
Strand=Plus/Plus

Query 3649197 GTTGTAAAGGTTAAGCCTCACGGATCATTAGTACTGGTTAGCTCAATACATCGCTGCACCT 3649256
          ||| || ||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
Sbjct 4557914 GTTATATGGTCAAGCCGACGGATCATTAGTATCAGTTAGCTCAATACATTGCTGTACTT 4557973

Query 3649257 ACACACCCAGCCTATCAACGTCATAGTCTTTAACGTTCCCTTACGGGGGCTTTAAGCCCCA 3649316
          ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
Sbjct 4557974 ACACACCTGACCTATCAACCACATAGTCTATATGGTTCCCTTACGGGGGCTT-GTGCCCC- 4558031

Query 3649317 GGGAAAGACTCATCTCGAGGCAAGTTTCCCGCTTAGATGCTTTACAGCGTTATCTCTCCG 3649376
          ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
Sbjct 4558032 GGGAAAGTCTCATCTGAGGCGCGCTTCCCGCTTAGATGCTTTACAGCGTTATCGCTCCG 4558091

Query 3649377 AATTTAGCTACCGGGCAATGCCATTGGCATGACAACCCGAACACAGTGATTGCTCCACT 3649436
          || ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
Sbjct 4558092 AACATAGCTACCCGGCAATGCCACTGGCGTGACAACCCGAACACAGAGGTTGCTCCACT 4558151

```

Figura 4 – Mostra os MISMATCHES, em amarelo, os MATCHS, em verde, e os GAPS em vermelho após a comparação de duas sequências. Fonte: Autor.

O BLAST+, entrega aos usuários uma gama de opções de comparação entre sequências, como ilustrado na Tabela 1, auxiliando cada vez mais os cientistas da área em sua finalidade.

A saída do BLAST+ fornece informações detalhadas sobre as sequências semelhantes encontradas, incluindo identidade, cobertura, posições de início e fim, e escore. Além disso, BLAST+ oferece suporte a diversos formatos de saída, incluindo tabela, formato XML e formato Alignment Viewer (AV).

2.5 Filtros

Ao realizar uma comparação de sequências de DNA ou proteínas através do algoritmo BLAST, é possível obter um grande número de resultados devido à alta sensibilidade do algoritmo. Embora esta sensibilidade alta seja confiável, ela também torna o processo de alinhamento mais lento. Para encontrar os alinhamentos mais significativos dentre todos os resultados, é necessário aplicar filtros.

A filtragem por E-Value é um método importante para reduzir a quantidade de alinhamentos gerados pelo BLAST, pois ele utiliza uma variável para limitar os alinhamentos de igual até menor valor dessa variável, a variável E-Value, também conhecida como

Tabela 1 – Variações do Blast+ e suas descrições.

Tipo de BLAST+	Descrição
blastn	Compara sequências de nucleotídeos com sequências de nucleotídeos.
blastp	Compara sequências de proteínas com sequências de proteínas.
blastx	Compara sequências de nucleotídeos traduzidas para proteínas com sequências de proteínas.
tblastn	Compara sequências de proteínas com sequências de nucleotídeos traduzidas
tblastx	Compara sequências de nucleotídeos traduzidas para proteínas com sequências de nucleotídeos traduzidas
psiblast	Utiliza uma sequência de proteínas como consulta para realizar uma busca por similaridade em uma base de dados de proteínas

"Expect". O E-Value é uma medida da probabilidade de ocorrência de um alinhamento acidental e, portanto, quanto menor o E-Value, maior é a confiabilidade do alinhamento.

A filtragem por identidade é outra forma de filtrar os alinhamentos gerados pelo BLAST. Aqui, é estabelecido uma variável que limita o valor do score de cada alinhamento. Os alinhamentos com scores mais baixos serão descartados do resultado final, pois são considerados menos confiáveis. A identidade é uma medida da semelhança entre as sequências alinhadas e, quanto maior a identidade, maior é a confiabilidade do alinhamento.

Por fim, a filtragem pela Razão entre o Tamanho do Alinhamento e o Tamanho do Gene (RAG) é uma maneira de filtrar os alinhamentos gerados pelo BLAST de acordo com o tamanho do alinhamento em relação ao tamanho do gene. É estabelecido uma variável que representa uma razão limite para a janela criada na comparação com o tamanho do alinhamento do gene. Os alinhamentos que não atenderem a essa razão limite serão descartados, pois são considerados menos confiáveis.

As filtragens dos alinhamentos gerados pelo BLAST são importantes para encontrar os alinhamentos mais significativos. Isso ajuda a evitar resultados falsos positivos e a obter resultados mais precisos.

2.6 Kablammo

O Kablammo¹ é um visualizador web de resultados de pesquisa BLAST, uma ferramenta fundamental na bioinformática que permite a identificação de homologies entre sequências genéticas ou proteicas. Desenvolvido com o objetivo de tornar a análise dos resultados do BLAST mais rápida e intuitiva, o Kablammo oferece recursos avançados, como visualização gráfica dos alinhamentos, visualização em árvore filogenética e mapeamento de domínios de proteínas, que permitem ao usuário uma melhor compreensão dos resultados obtidos.

Além disso, a ferramenta permite a importação de sequências personalizadas para comparação e o download dos resultados em diferentes formatos. Como um projeto de código aberto, o Kablammo também oferece a flexibilidade de ser personalizado e adaptado às necessidades específicas dos usuários. É, portanto, uma opção popular e valiosa para pesquisadores que precisam analisar e interpretar grandes quantidades de dados de sequenciamento genético ou proteico, como ilustrado na Figura 5.

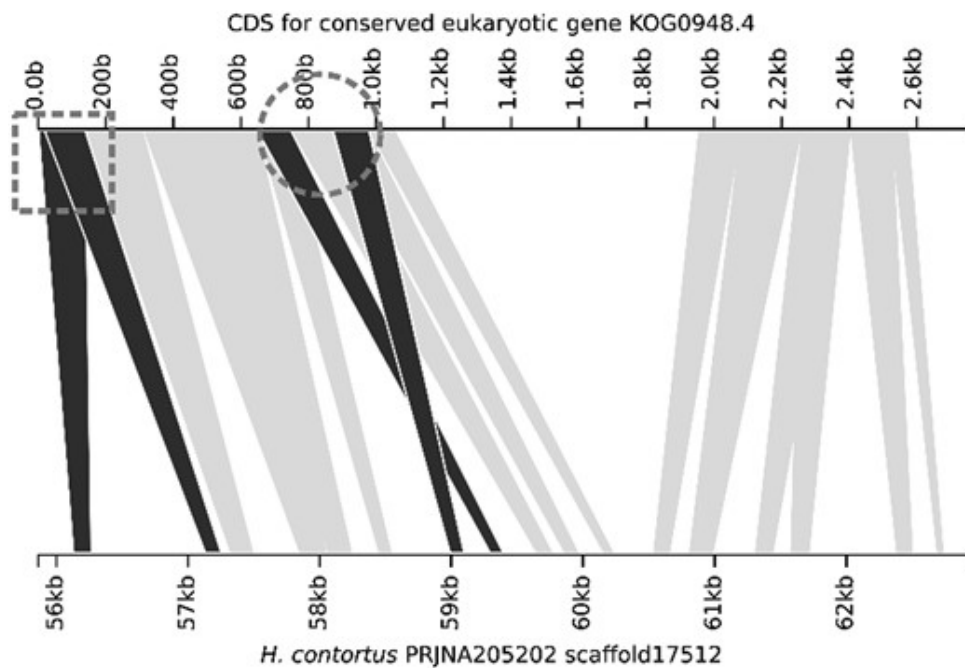


Figura 5 – Usando Kablammo para investigar a estrutura do gene em uma montagem do genoma de *Haemonchus contortus*. Fonte: (WINTERSINGER; WASMUTH, 2014)

¹ <https://github.com/jwintersinger/kablammo>

3 Trabalhos Relacionados

Esta seção apresenta trabalhos relevantes e relacionados a este projeto. A Tabela 2 mostra uma compilação desses trabalhos, categorizados de acordo com três métricas importantes: plataforma, objetivo e tecnologias.

A métrica Plataforma se refere à plataforma utilizada na aplicação descrita em cada trabalho. Por exemplo, o trabalho de [Mueller et al. \(2008\)](#) apresenta um visualizador comparativo baseado na web, projetado para mapeamento de dados genéticos, físicos e citológicos. Já o trabalho de [Youens-Clark et al. \(2009\)](#) apresenta uma ferramenta na web para exibição e comparação de mapas. O trabalho de [Batesole et al. \(2014\)](#) descreve o desenvolvimento de uma plataforma na web chamada YouGenMap, utilizada para mapeamento e visualização simples de mapas. O trabalho de [Overmars et al. \(2015\)](#) apresenta uma ferramenta na web para visualização circular de dados. Por fim, o trabalho de [Holtz, David e Ranwez \(2017\)](#) descreve o desenvolvimento da aplicação Genetic Map Comparator, uma ferramenta na web para comparação de mapas genéticos.

A métrica Objetivo destaca os objetivos principais de cada um dos trabalhos relacionados a este projeto. Por exemplo, o trabalho de [Mueller et al. \(2008\)](#) oferece uma plataforma web que permite aos usuários fazer upload de seus próprios mapas e compará-los com outros mapas no sistema. O trabalho de [Youens-Clark et al. \(2009\)](#) é uma ferramenta web que permite ao usuário comparar vários mapas e exibir a comparação resultante. O trabalho de [Batesole et al. \(2014\)](#) é uma ferramenta web de código aberto que oferece uma funcionalidade de comparação dinâmica de mapas baseada nas entradas dos usuários, permitindo comparar dois ou mais conjuntos de mapas. O trabalho de [Overmars et al. \(2015\)](#) apresenta uma ferramenta web que gera mapas circulares personalizados para ajudar na análise de genomas bacterianos e elementos de sequência. O trabalho de [Holtz, David e Ranwez \(2017\)](#) apresenta uma aplicação web de comparação de mapas que se baseia nas estatísticas e nas posições relativas de marcadores comuns.

Por fim, a métrica Tecnologias destaca as ferramentas e técnicas utilizadas pelos trabalhos relacionados. Cada trabalho mencionado utiliza tecnologias distintas, por exemplo: o trabalho de [Mueller et al. \(2008\)](#) utiliza Perl orientado a objetos e um banco de dados, enquanto o trabalho de [Youens-Clark et al. \(2009\)](#) foi implementado na linguagem Perl e executado em um servidor Web Apache. O trabalho de [Batesole et al. \(2014\)](#) foi desenvolvido através da combinação de várias bibliotecas JavaScript e PHP. Já o trabalho de [Overmars et al. \(2015\)](#) utilizou a linguagem Python, Javascript, SVG e um banco de dados MySQL. Por fim, o trabalho de [Holtz, David e Ranwez \(2017\)](#) foi implementado com R Shiny e Plotly.

Em resumo, a Tabela 2 apresenta uma comparação dos trabalhos relacionados a esse projeto, ajudando a entender as diferenças e semelhanças entre eles, e fornecendo uma

base para a construção deste projeto.

Tabela 2 – Trabalhos Relacionados.

Trabalho	Plataforma	Objetivo	Tecnologias
(MUELLER et al., 2008)	Aplicação Web	Visualizador comparativo.	Perl Orientado a Objetos.
(YOUENS-CLARK et al., 2009)	Aplicação Web	Uma ferramenta para exibir e comparar mapas.	Perl, MySQL, HTML e JavaScript
(BATESOLE et al., 2014)	Aplicação Web	Uma ferramenta para mapeamento comparativo dinâmico de mapas.	JavaScript, HTML, PHP e MySQL
(OVERMARS et al., 2015)	Aplicação Web	Uma ferramenta que cria mapas circulares para análise.	Python, Javascript, MySQL e SVG.
(HOLTZ; DAVID; RANWEZ, 2017)	Aplicação Web	Uma aplicação para comparação de mapas	R shiny e Plotly

Os trabalhos relacionados apresentados fornecem uma visão geral das ferramentas e técnicas utilizadas na comparação de mapas genéticos. Existem diversas ferramentas disponíveis para a comparação de mapas genéticos, cada uma com suas vantagens e limitações. O Mueller et al. (2008) é um visualizador comparativo simples, sem algoritmos de comparação. Já o Youens-Clark et al. (2009) é um visualizador e comparador de mapas que utiliza a ferramenta CMAP, mas não realiza alinhamento genético. O Batesole et al. (2014) possui um visualizador com poucas informações sobre os alinhamentos em comparação. O Overmars et al. (2015) trabalha com um visualizador de mapas circulares e não possui comparação genética. O Holtz, David e Ranwez (2017) é uma boa ferramenta web para comparação de mapas genéticos, mas não tem visualizador gráfico dos resultados.

A ferramenta GeMapCom difere dos trabalhos citados no capítulo por seu objetivo principal, que é fornecer uma plataforma desktop para realização de alinhamentos utilizando diferentes algoritmos, tais como o Needleman-Wunsch, Smith-Waterman e BLAST. Além disso, a ferramenta permite salvar os resultados gerados em diferentes formatos, como imagens ou textos bem estruturados, e apresentar esses resultados de forma dinâmica, seja por meio de gráficos interativos ou visualizações detalhadas.

Sendo válido ressaltar que a utilização de um aplicativo desktop, como no caso do BLAST+ executando localmente, permite ao usuário um controle mais detalhado sobre as configurações de comparação, o que pode não ser possível em uma versão web do BLAST. Isso ocorre porque a versão web pode ter algumas limitações de personalização devido a preocupações com a segurança e a estabilidade do servidor. Além disso, a versão

web pode impor limitações no tamanho do arquivo de entrada ou no tempo de execução, que podem ser superadas em um aplicativo desktop executado localmente. Dessa forma, a utilização do BLAST+ em um aplicativo desktop permite que o usuário personalize o processo de acordo com suas necessidades específicas e realize análises mais complexas e extensas sem limitações.

Diferentemente de outras ferramentas, o GeMapCom não se restringe a comparações genéticas, permitindo a comparação de diversas formas possíveis, como por exemplo comparações de sequências de aminoácidos. Além disso, a ferramenta evita a necessidade de pesquisar em diferentes bancos de dados na internet em busca de um sequenciamento específico, pois permite a realização dessas buscas diretamente na aplicação.

4 Metodologia Proposta

Este capítulo apresenta informações sobre a ferramenta GeMapCom¹, abordando a metodologia seguida e a lógica do seu funcionamento.

4.1 Diagrama de caso de uso

O diagrama de casos de uso resume os detalhes dos usuários do sistema e as interações deles com o sistema, definido na *Unified Modeling Language* (UML), que tem como objetivo a elaboração da estrutura de projetos de software.

A Figura 6 demonstra o ator, definido pela UML como o usuário do sistema, interagindo com a ferramenta GeMapCom, além de mostrar os casos de uso, que são tarefas ou funcionalidades realizadas pelo ator.

4.2 Ferramentas e Tecnologias

As ferramentas e tecnologias utilizadas neste trabalho foram selecionadas com base na eficiência e na adequação às necessidades do projeto. Foram utilizadas ferramentas amplamente reconhecidas pela comunidade científica e tecnológica, que proporcionam segurança e confiabilidade ao processo de análise de dados genômicos.

O *python* é uma linguagem de programação de fácil aprendizado e de alto desempenho, tornando-a ideal para desenvolvimento de aplicativos, especialmente em ciência de dados e biologia molecular. Além disso, o *Python* possui uma ampla comunidade e diversas bibliotecas prontas para serem utilizadas, o que facilita o desenvolvimento de aplicativos complexos.

A *BioPython* é uma biblioteca de software livre para biologia computacional em *Python*. Ela fornece ferramentas para manipulação de dados biológicos, tais como sequências de DNA, proteínas e estruturas tridimensionais, bem como realizar tarefas comuns na bioinformática, como alinhamentos múltiplos, análise de sequências, busca de padrões e muito mais. A *BioPython* é amplamente utilizada por cientistas da computação e biólogos devido a sua capacidade de integração com outras bibliotecas, a grande quantidade de recursos disponíveis.

O *Qt* é uma ferramenta de desenvolvimento de interfaces gráficas de usuários, que permite a criação de aplicativos com aparência profissional e com suporte a multiplataformas. Além disso, o *Qt* possui recursos avançados como modelos de dados, suporte a

¹ <https://github.com/PedroAzevedo141/GeMapCom>

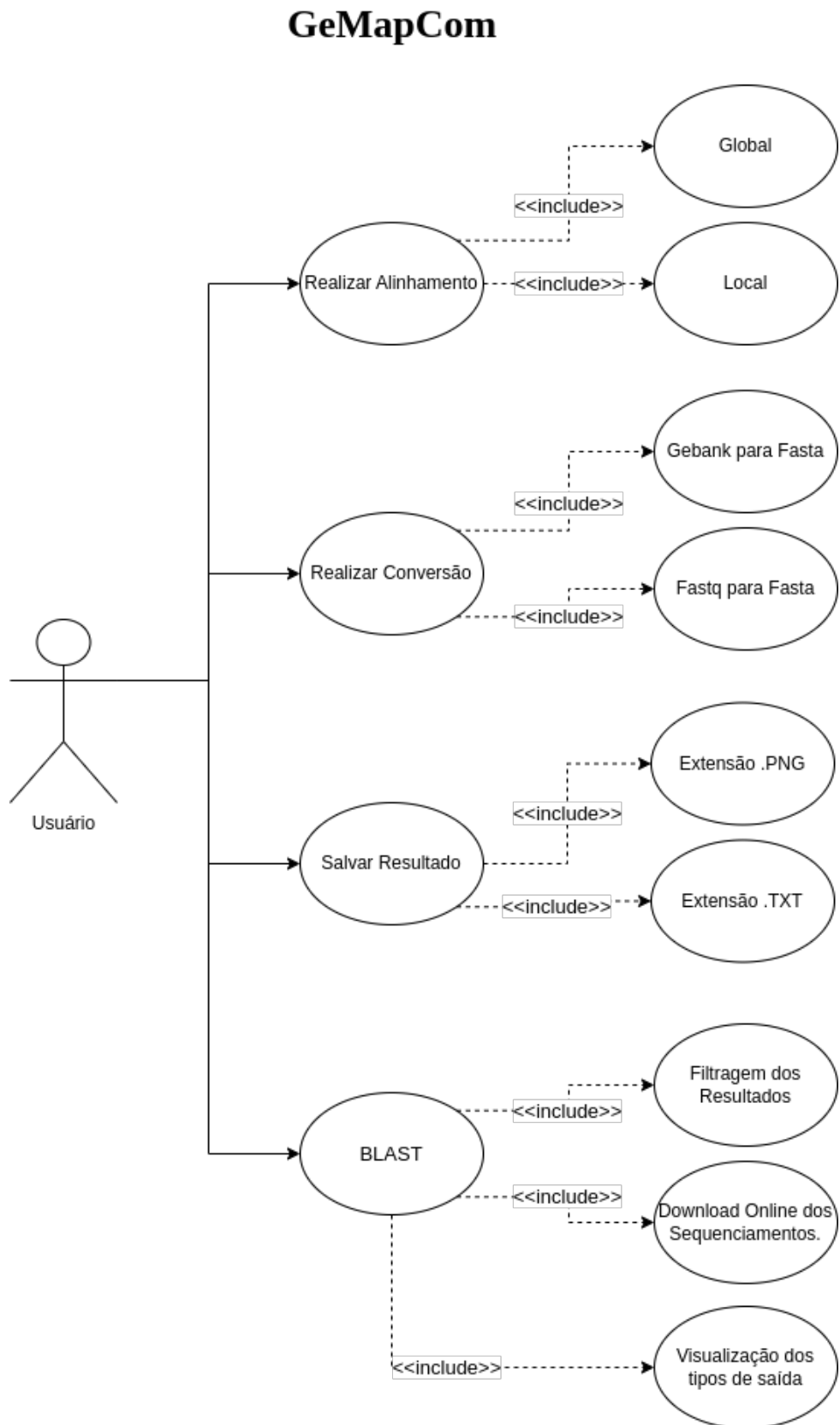


Figura 6 – Diagrama de Caso de Uso com o Ator Interagindo com a Ferramenta GeMapCom. Fonte: Autor.

diversos tipos de arquivos e interação com servidores, tornando-o uma escolha ideal para o desenvolvimento de aplicativos desktop.

4.3 Lógica de Desenvolvimento

A lógica utilizada para o desenvolvimento deste trabalho teve como conceito principal tornar a experiência do usuário o mais satisfatória possível, proporcionando uma interação intuitiva com a ferramenta e apresentando as informações de forma clara e coerente. A arquitetura utilizada foi planejada para facilitar a comunicação entre o usuário e a ferramenta. A Figura 7 fornece uma visão clara da arquitetura utilizada.

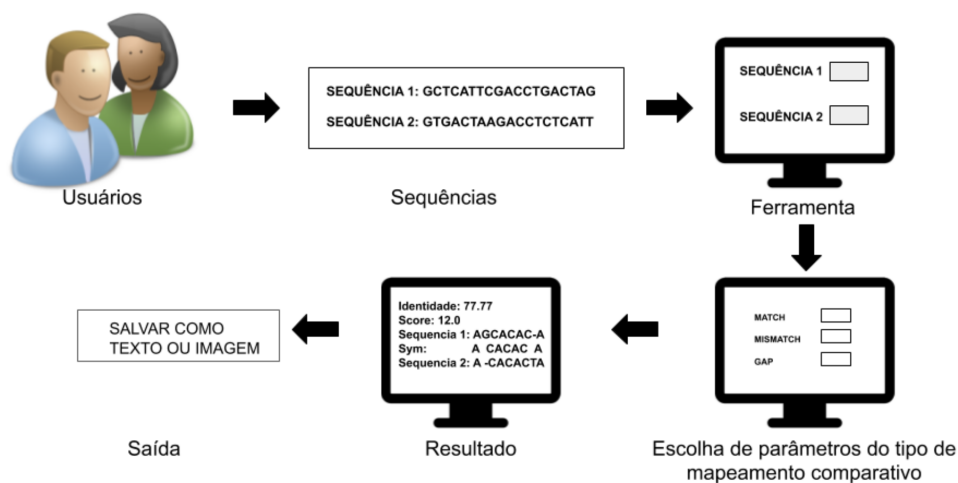


Figura 7 – Arquitetura da Ferramenta GeMapCom. Fonte: Autor.

Ao utilizar a ferramenta, o usuário pode esperar ter acesso a informações claras e objetivas, que lhe permitirão tomar decisões mais informadas e assertivas. Além disso, a interação intuitiva com a ferramenta torna o processo mais simples e eficiente, otimizando o tempo e esforço investido pelo usuário. O desenvolvimento desta metodologia busca atender às demandas e necessidades dos usuários, a fim de garantir que a ferramenta seja utilizada de forma eficaz e satisfatória.

5 Resultados e Discussões

Esta seção tem como objetivo apresentar as seis telas principais da ferramenta, juntamente com seus respectivos *pop-up*. A ferramenta inclui a Tela Inicial, a Tela de Comparação Local e Global, a Tela de Conversão de Arquivo .gbk para .fasta, a Tela do Blast+, a Tela de Filtro dos Resultados do Blast+ e a Tela de Resultados, que contém informações em formato de caracteres e gráficos.

5.1 Tela Inicial

A Figura 8 mostra a tela principal da ferramenta, com uma interface disponibilizando as seguintes funções, que podem ser utilizadas pelos usuários: Comparar Sequências; Converter Sequência; Blast+; Outras Ferramentas; e Sair.

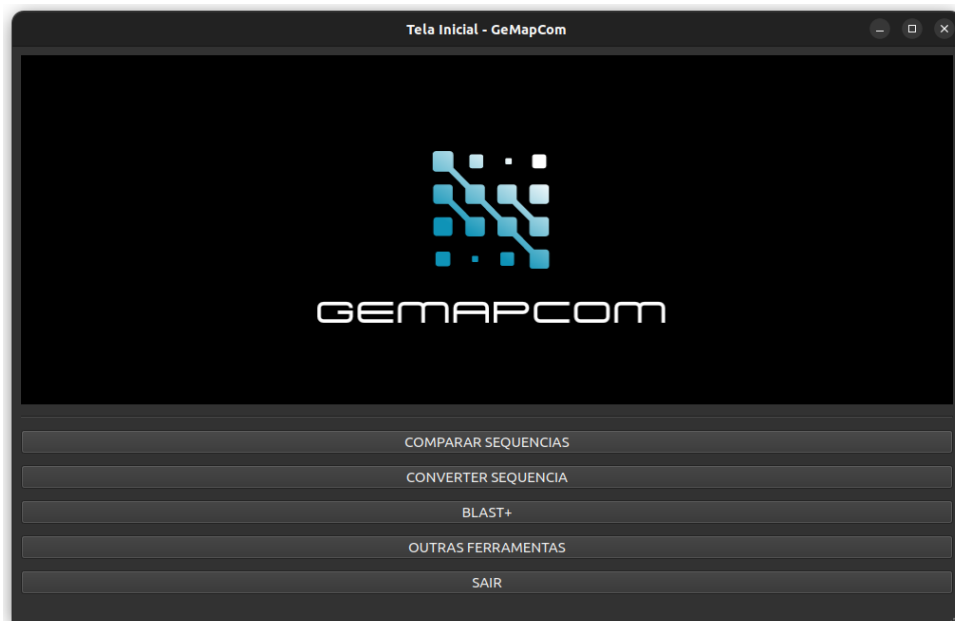


Figura 8 – Tela Principal da Ferramenta GeMapCom. Fonte: Autor.

5.2 Tela de Comparar Sequências

A tela de comparação, ilustrada na Figura 9, é uma das funcionalidades da ferramenta GeMapCom que permite aos usuários comparar sequências de DNA ou proteína. Nessa tela, os usuários podem escolher entre dois tipos de comparação: global e local, utilizando os algoritmos Needleman-Wunsch e Smith-Waterman, respectivamente. Essa opção de escolha entre diferentes algoritmos possibilita a adaptação da análise às necessidades

específicas de cada usuário, levando em conta fatores como o tamanho das sequências e a complexidade do alinhamento.

Além disso, na tela de comparação, os usuários têm a possibilidade de definir os parâmetros de alinhamento das sequências, como os valores de match, mismatch e gap. Esses parâmetros podem ser ajustados para que o alinhamento resultante atenda às necessidades específicas do usuário em termos de sensibilidade e especificidade, por exemplo.

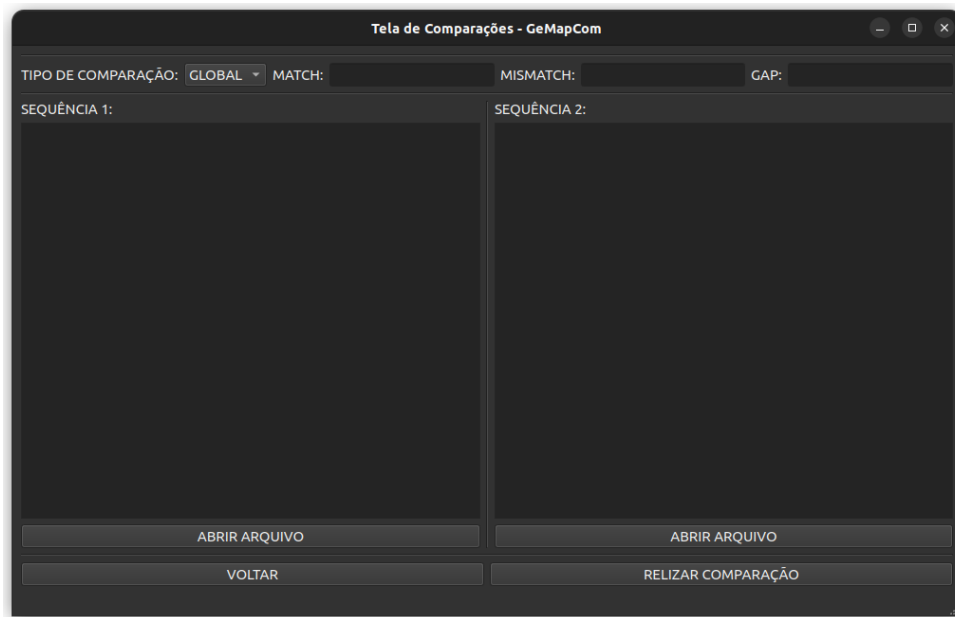


Figura 9 – Tela de Comparação da GeMapCom. Fonte: Autor.

5.3 Tela de Conversão

A Figura 10 ilustra a tela de conversão de formatos da ferramenta, que permite aos usuários converter arquivos de sequências em dois dos principais formatos utilizados na bioinformática: Genbank e Fasta, além de converter arquivos Fastq para Fasta. Essa funcionalidade é crucial para garantir a interoperabilidade entre diferentes ferramentas e bancos de dados, permitindo que os pesquisadores possam trabalhar com uma variedade de dados de sequenciamento genético, independentemente do formato em que estão disponíveis.

A conversão de formatos de arquivos de sequências é uma etapa fundamental em muitos fluxos de trabalho na bioinformática. No caso da ferramenta GeMapCom, a conversão de arquivos para o formato Fasta é especialmente importante, pois a ferramenta só é capaz de processar arquivos nesse formato. Dessa forma, a conversão para o formato Fasta é uma etapa necessária para o uso da ferramenta.



Figura 10 – Tela de Conversão da GeMapCom. Fonte: Autor.

5.4 Tela do Blast+

A Figura 11 mostra a tela do Blast, onde o usuário pode interagir com o software da NCBI localmente em sua máquina. É possível informar os campos de match, mismatch e gap, e escolher o formato de saída desejado. Além disso, a tela também possui a opção de buscar sequências online em bancos de dados genômicos e escolher o tipo de alinhamento desejado por meio de um pop-up ilustrado na Figura 12. Mais detalhes sobre as aplicações do Blast estão descritos na Tabela 1.

O pop-up ilustrado na Figura 13 é aberto ao clicar no botão de pesquisa na tela do BLAST e permite buscar sequências online nos bancos de dados genômicos. Nessa tela, o usuário pode inserir um endereço de e-mail opcional para a requisição ser direcionada para a conta informada no banco de dados, escolher o banco de dados em que a sequência será pesquisada e inserir o ID de armazenamento da sequência desejada. Além disso, no canto inferior da imagem, há um campo que mostra a sequência retornada para uma pré-visualização antes de retornar para a tela do BLAST novamente. Essa funcionalidade permite que o usuário acesse uma ampla gama de sequências disponíveis nos bancos de dados genômicos e facilite a realização de análises de sequências mais abrangentes e precisas.

5.5 Tela de Filtros

A Figura 14 mostra a tela responsável por realizar a filtragem dos dados dos alinhamentos resultantes do BLAST. Nessa tela, o usuário pode aplicar três tipos de filtros diferentes: filtro de identidade, filtro de valor E e filtro de razão entre o tamanho do ali-

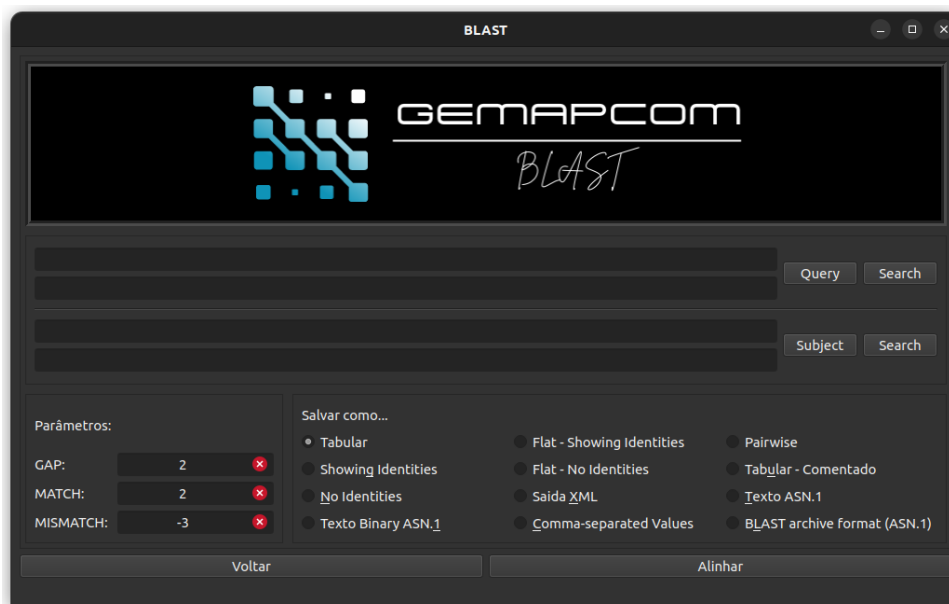


Figura 11 – Tela principal do Blast. Fonte: Autor.

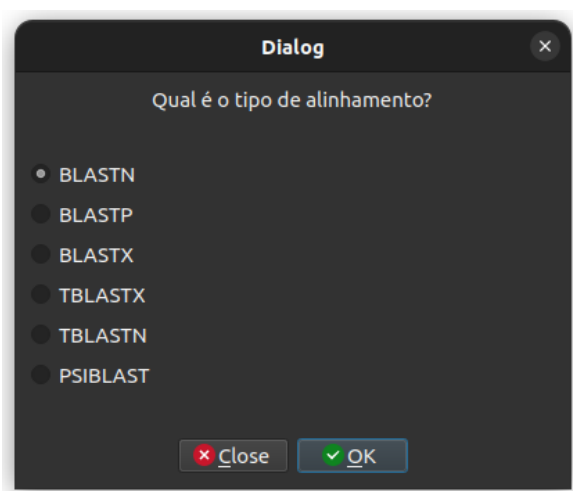


Figura 12 – Tela de seleção de tipo de alinhamento do Blast. Fonte: Autor.

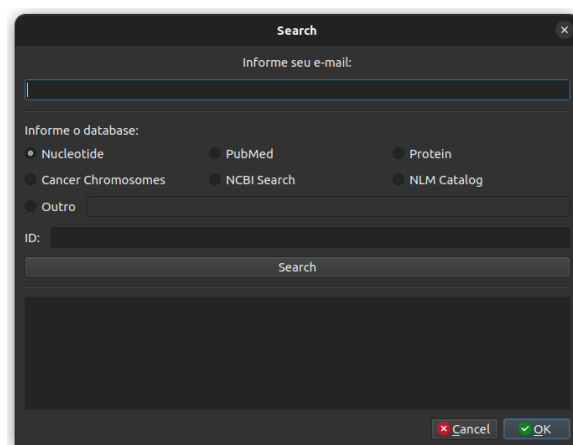


Figura 13 – Tela de busca de seqüências online da ferramenta GeMapCom. Fonte: Autor.

nhamento e o tamanho do gene (RAG), como descrito na seção 2.5. Esses filtros permitem ao usuário personalizar a precisão e a abrangência dos resultados do BLAST, selecionando apenas os alinhamentos que atendem aos seus critérios de qualidade e relevância.

5.6 Tela de Resultados

A tela de resultados apresenta a representação dos resultados obtidos nos alinhamentos, que podem ser apresentados em forma escrita ou gráfica. A Figura 15 ilustra um exemplo de resultado obtido após um alinhamento realizado pelo BLAST, exibindo o alinhamento, os matches, os mismatches e os gaps, bem como a porcentagem de identidade e o score de cada alinhamento resultante.



Figura 14 – Tela de Filtragem do Blast. Fonte: Autor.

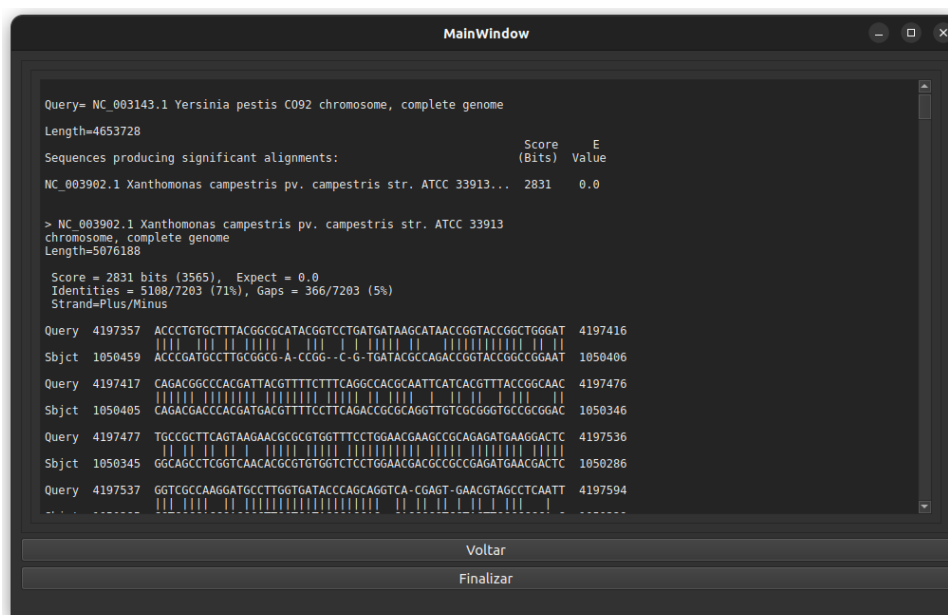


Figura 15 – Tela de Resultados da GeMapCom. Fonte: Autor.

Ao escolher o formato de saída em XML para o alinhamento, a ferramenta permite visualizar cada alinhamento resultante em gráficos pela ferramenta de visualização open-source Kablamm. Sendo possível interagir com os gráficos de maneira dinâmica e interativa, como mostrado na Figura 16. Ao passar o mouse sobre os alinhamentos, as informações exibidas à direita do gráfico são atualizadas instantaneamente, proporcionando uma experiência interativa e envolvente para o usuário.

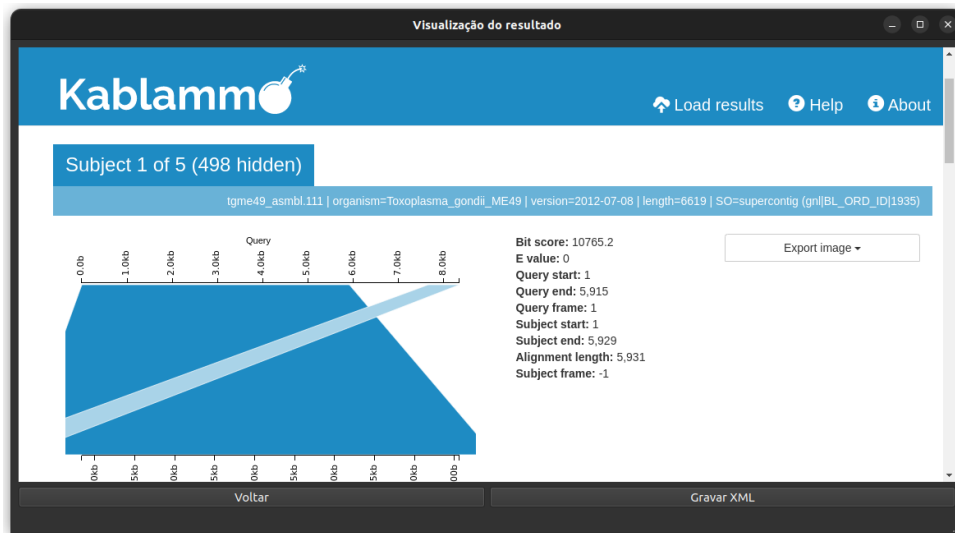


Figura 16 – Tela de Resultados Gráficos do Blast. Fonte: Autor.

Além disso, na Figura 16, é possível observar que é possível selecionar um alinhamento de interesse no gráfico e visualizar seu alinhamento destacado por cores. Essa opção permite que o usuário se concentre em um alinhamento específico e o analise com mais detalhes, o que pode ser muito útil em muitas aplicações, como a pesquisa de homologies entre sequências biológicas. A Figura 17 ilustra esse recurso.

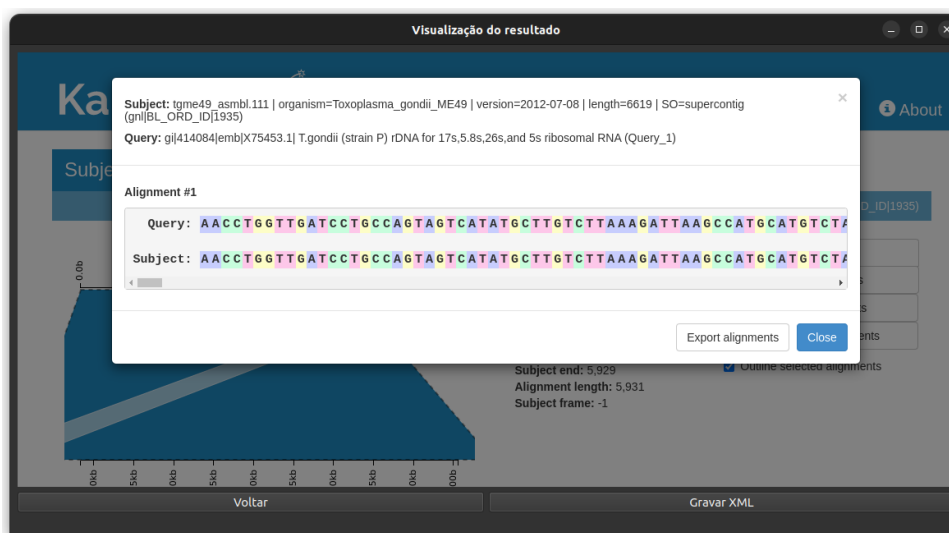


Figura 17 – Tela de Resultados Gráficos do Blast dando ênfase a possível interação do usuário com os gráficos. Fonte: Autor.

6 Conclusão

Este trabalho apresentou uma ferramenta desktop que facilita a interação com as funcionalidades do Blast+ e fornece diversas formas de visualização dos resultados, apresentando informações claras e organizadas. Tendo todos os requisitos definidos para a execução do projeto alcançados, resultando em um registro de software no Instituto Nacional de Propriedade Industrial (INPI), como ilustrado no Apêndice A.

Com a criação desta ferramenta, espera-se que seja possível contribuir para o avanço da pesquisa na área, permitindo aos pesquisadores uma análise mais clara e precisa de dados. A ferramenta desenvolvida é uma solução prática e eficiente para a análise dos sequenciamentos e apresenta potencial para futuras melhorias e aprimoramentos.

Uma das possibilidades para trabalhos futuros é realizar uma avaliação de usabilidade da ferramenta, a fim de garantir que ela atenda às necessidades dos usuários de forma eficiente e intuitiva. Com base nos resultados da avaliação, poderão ser identificados pontos de melhoria e ajustes necessários na interface e nas funcionalidades da ferramenta, aprimorando sua acessibilidade e eficácia para os usuários. Essa avaliação poderá fornecer informações valiosas para o desenvolvimento contínuo da ferramenta e para melhorar sua contribuição no campo da bioinformática.

Referências

- ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, Elsevier, v. 215, n. 3, p. 403–410, 1990. Citado na página 18.
- BACHHAWAT, A. K. Comparative genomics: A powerful new tool in biology. *Resonance*, Springer, v. 11, p. 22–40, 2006. Citado na página 15.
- BATESOLE, K. et al. Yougenmap: a web platform for dynamic multi-comparative mapping and visualization of genetic maps. *Frontiers in genetics*, Frontiers, v. 5, p. 183, 2014. Citado 2 vezes nas páginas 22 e 23.
- BENSON, D. A. et al. Genbank. *Nucleic acids research*, Oxford University Press, v. 43, n. Database issue, p. D30, 2015. Citado na página 12.
- CHENNA, R. et al. Multiple sequence alignment with the clustal series of programs. *Nucleic acids research*, Oxford University Press, v. 31, n. 13, p. 3497–3500, 2003. Citado na página 18.
- DOOLITTLE, R. F. Molecular evolution: computer analysis of protein and nucleic acid sequences. *Methods in enzymology*, v. 183, 1990. Citado na página 16.
- ENA. *European Nucleotide Archive*. [S.l.], 2020. Disponível em: <<https://www.ebi.ac.uk/ena/browser/home>>. Acesso em: 29 ago. 2020. Citado na página 16.
- GOLUBCHIK, T. et al. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Molecular biology and evolution*, Society for Molecular Biology and Evolution, v. 24, n. 11, p. 2433–2442, 2007. Citado na página 17.
- GOTOH, O. An improved algorithm for matching biological sequences. *Journal of molecular biology*, Elsevier, v. 162, n. 3, p. 705–708, 1982. Citado na página 17.
- HARDISON, R. C. Comparative genomics. *PLoS biology*, Public Library of Science San Francisco, USA, v. 1, n. 2, p. e58, 2003. Citado na página 15.
- HOLTZ, Y.; DAVID, J. L.; RANWEZ, V. The genetic map comparator: a user-friendly application to display and compare genetic maps. *Bioinformatics*, v. 33, n. 9, 2017. Citado 2 vezes nas páginas 22 e 23.
- ISAEV, A. et al. *Introduction to mathematical methods in bioinformatics*. [S.l.]: Springer, 2004. Citado na página 17.
- JGI. *Integrated Microbial Genomes System*. [S.l.], 2018. Disponível em: <<https://img.jgi.doe.gov/>>. Acesso em: 29 ago. 2020. Citado na página 16.
- JUNIOR, N. F. de A. Ferramentas para comparação genômica. 2002. Citado na página 12.
- KIM, S. et al. PubChem 2023 update. *Nucleic Acids Research*, v. 51, n. D1, p. D1373–D1380, 10 2022. ISSN 0305-1048. Disponível em: <<https://doi.org/10.1093/nar/gkac956>>. Citado na página 12.

MUELLER, L. A. et al. The sgn comparative map viewer. *Bioinformatics*, Oxford University Press, v. 24, n. 3, p. 422–423, 2008. Citado 2 vezes nas páginas 22 e 23.

National Human Genome Research Institute. *Comparative Genomics Fact Sheet*. 2020. <https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet>. Citado na página 15.

National Human Genome Research Institute. *Online Education Kit: Bioinformatics Introduction*. 2021. <https://www.genome.gov/25020000/online-education-kit-bioinformatics-introduction>. Citado na página 12.

NCBI. *National Center for Biotechnology Information*. [S.l.], 2018. Disponível em: <<https://www.ncbi.nlm.nih.gov/>>. Acesso em: 29 ago. 2020. Citado na página 16.

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*, Elsevier, v. 48, n. 3, p. 443–453, 1970. Citado na página 17.

OVERMARS, L. et al. Civi: circular genome visualization with unique features to analyze sequence elements. *Bioinformatics*, Oxford University Press, v. 31, n. 17, p. 2867–2869, 2015. Citado 2 vezes nas páginas 22 e 23.

PIR. *Protein Information Resource*. [S.l.], 2020. Disponível em: <<https://proteininformationresource.org/>>. Acesso em: 29 ago. 2020. Citado na página 16.

SIVASHANKARI, S.; SHANMUGHAVEL, P. Comparative genomics-a perspective. *Bioinformatics*, Biomedical informatics publishing Group, v. 1, n. 9, p. 376, 2007. Citado na página 15.

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. Citado na página 17.

WANG, B. *Implementation of a dynamic programming algorithm for DNA sequence alignment on the Cell MatrixTM architecture*. [S.l.]: Utah State University, 2002. Citado na página 17.

WINTERSINGER, J. A.; WASMUTH, J. D. Kablammo: an interactive, web-based BLAST results visualizer. *Bioinformatics*, v. 31, n. 8, p. 1305–1306, 12 2014. ISSN 1367-4803. Disponível em: <<https://doi.org/10.1093/bioinformatics/btu808>>. Citado 2 vezes nas páginas 8 e 21.

YOUENS-CLARK, K. et al. Cmap 1.01: a comparative mapping application for the internet. *Bioinformatics*, Oxford University Press, v. 25, n. 22, p. 3040–3042, 2009. Citado 2 vezes nas páginas 22 e 23.

APÊNDICE A – Registro de Software



REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA ECONOMIA
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512021001837-0**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 01/03/2021, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

Título: GeMapCom: Uma ferramenta para mapeamento comparativo e visualização de dados Genômicos.

Data de criação: 01/03/2021

Titular(es): UNIVERSIDADE FEDERAL DO PIAUÍ

Autor(es): ROMUERE RODRIGUES VELOSO E SILVA; ANTONIO OSEAS DE CARVALHO FILHO; FLAVIO HENRIQUE DUARTE DE ARAUJO; FRANCISCO DAS CHAGAS DOS ANJOS CARVALHO JUNIOR; PEDRO AZEVEDO ABRANTES DE OLIVEIRA; LILIAN ROSALINA GOMES SILVA; JOSÉ LINDEMBERG ROCHA SARMENTO

Linguagem: PYTHON

Campo de aplicação: BL-02

Tipo de programa: AP-01; DS-04

Algoritmo hash: SHA-512

Resumo digital hash:

47835E441AE6CE5D3824517D4B25D394F9BB4800D13FC32A865DC11EF5E9B2AB3DDE7349A06E7EA94A342111
8934A370004A951BA20A492FA98CB8F629391A9A

Expedido em: 10/08/2021

Aprovado por:

Carlos Alexandre Fernandes Silva
Chefe da DIPTO

APÊNDICE B – Certificado

CERTIFICADO

Certificamos que **Pedro Azevedo Abrantes de Oliveira** (discente, PIBITI CNPq (IT)) e **Prof. Dr. Romuere Rodrigues Veloso e Silva** (orientador(a), UNIVERSIDADE FEDERAL DO PIAUÍ) obtiveram o **1.º Lugar** em **TECNOLOGIAS DE PRODUÇÃO** na apresentação oral do trabalho intitulado **UMA FERRAMENTA PARA MAPEAMENTO COMPARATIVO E VISUALIZAÇÃO DE DADOS GENÔMICOS**, durante o **XIII SDTI: XIII Seminário de Desenvolvimento Tecnológico e Inovação - 2021**, realizado na UFPI – Campus Ministro Petrônio Portella – Teresina – PI.



Prof. Dr. Gildásio Guedes Fernandes
Reitor da Universidade Federal do Piauí
Presidente do SIUFPI



Prof. Dra. Deborah Dettman Matos
Pró-Reitora de Extensão e Cultura - PREXC
Coordenadora Geral do SIUFPI



Prof. Dra. Keylla Maria de Sá Urtiga Aita
Coord. de Pesquisa e Inovação PROPESQI/UFPI
Coordenadora XIII SDTI



TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA “JOSÉ ALBANO DE MACEDO”

Identificação do Tipo de Documento

- Tese
- Dissertação
- Monografia
- Artigo

Eu, Pedro Azevedo Abrantes de Oliveira, autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação **“GeMapCom 2.0: Uma Ferramenta para Mapeamento Comparativo e Visualização Interativa de Dados Genômicas”** de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI, 24 de Março de 2023.

Documento assinado digitalmente
gov.br PEDRO AZEVEDO ABRANTES DE OLIVEIRA
Data: 24/03/2023 18:03:55-0300
Verifique em <https://validar.iti.gov.br>

Assinatura