UNIVERSIDADE FEDERAL DO RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação (Mestrado)

ESSEX: Identificação de um aminoácido de interesse em sequências biológicas de origens diferentes

Wolmer Dias Quaresma Junior

Dissertação apresentado ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Computação

Orientador: Prof. Dr. Adriano Velasque Werhli
Co-orientador: Prof. Dr. Karina dos Santos Machado

Rio Grande, 2019

Ficha catalográfica

Q18e Quaresma Junior, Wolmer Dias.
Essex : identificação de um aminoácido de interesse em sequências biológicas de origens diferentes / Wolmer Dias Quaresma Junior. – 2019.
96 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Computação, Rio Grande/RS, 2019.
Orientador: Dr. Adriano Velasque Werhli.
Coorientadora: Dra. Karina dos Santos Machado.

1. Sequência biológica 2. Alinhamento global 3. Alinhamento local 4. Alinhamento múltiplo 5. Essex I. Werhli, Adriano Velasque II. Machado, Karina dos Santos III. Título.

CDU 004:577

Catalogação na Fonte: Bibliotecária Vanessa Ceiglinski Nunes CRB 10/2174

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

DISSERTAÇÃO DE MESTRADO

ESSEX: Identificação de um aminoácido de interesse em sequências biológicas de origens diferentes

Wolmer Dias Quaresma Junior

Banca examinadora:

Diana F. Adamatti

Profª. Drª Diana Francisca Adamatti (FURG)

Profª. Drª Ana Trindade Winck (UFCSPA)

Adriano V. Werhli

Prof. Dr. Adriano Velasque Werhli
Orientador(a)

Karina Machado

Profª. Drª. Karina dos Santos Machado
Coorientador(a)

Dedico esse trabalho aos meus orientadores
Prof. Dr. Adriano Werhli e Prof^{a}. Dr^{a}. Karina Machado
pela paciência, confiança, amizade, incentivo e orientação.

AGRADECIMENTOS

A presente dissertação de mestrado não poderia chegar nesse momento sem o precioso apoio de diversas pessoas. Em primeiro lugar, não posso deixar de agradecer aos meus orientadores, Professor Adriano Werhli e a Professora Karina Machado, pelas discussões, paciência, pelos conhecimentos transmitidos e por todo empenho. E o mais importante, muito obrigado por me ter corrigido quando necessário sem nunca me desmotivar. Desejo igualmente agradecer a todos os meus colegas do Mestrado de Engenharia da Computação e em extensão ao grupo de Biologia Computacional, especialmente aos colegas Eduardo Abreu pelo jeito gaudério que sempre se mostrou preocupado, que me ligava dia e noite mostrando atenção e preocupação. Ao colega João Scaini, que sempre mostrou estar disposto ajudar a cada momento, a cada dica muito valiosa e apoio constantes. Ao colega Bruno Oliveira, que sempre me auxiliou a interpretar dados biológicos, ajudou nos textos e nas piadas internas do grupo Combi-Lab.

Aos membros da banca examinadora, Professora Diana Adamatti, agradeço pelas conversas breves, porém importantíssimas e motivacionais. A professora Ana Winck, que tão gentilmente aceitou participar e colaborar com esta dissertação. Agradeço aos funcionários do C3, que sempre foram prestáveis e me ajudaram a ultrapassar grandes obstáculos.

À minha mãe e ao meu pai deixo um agradecimento especial, por todas as lições de amor, companheirismo, amizade, caridade, dedicação, compreensão e oportunidade de viver com vocês a cada novo dia. Sinto-me orgulhoso e privilegiado por ter pais tão especiais. A minha irmã querida, por todas as conversas e sempre pronta a me apoiar em tudo nessa vida. Ao meu irmão, que tem seu jeito quieto, sempre demonstrou apoio e

admiração por minha trajetória. Ao meu cunhado, afiliado e sobrinha por cada sorriso no rosto que se tornou o combustível para poder continuar a caminhada.

A minha namorada e família, por todo amor incondicional que você sempre me deu. Que independente das circunstancias, você me fez acreditar que eu era capaz em tudo que eu fosse fazer. Obrigado por todo carinho, compreensão e apoio em tantos momentos difíceis desta caminhada. Obrigado pelo presente de cada dia, pelo seu amor e por saber me motivar!

Agradeço e dedico este mestrado PLENAMENTE a DEUS! Tudo que possuo na vida foi recebido com a tua bênção e por isso te agradeço todos os dias. Muito obrigado!

Por fim, a todos aqueles que contribuíram, direta ou indiretamente, para a realização desta dissertação, o meu sincero agradecimento.

A persistência é o caminho do êxito.
— CHARLES CHAPLIN

RESUMO

JUNIOR, Wolmer Dias Quaresma. **ESSEX: Identificação de um aminoácido de interesse em sequências biológicas de origens diferentes**. 2019. 97 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

O objetivo desta dissertação é propor uma ferramenta para tratar um problema importante da biologia computacional, que corresponde na localização de um determinado aminoácido em uma sequência biológica para que possa obter algum significado biológico. Obter essa informação muitas vezes é um processo complexo, pois a numeração deste aminoácido na sequência onde o experimento foi realizado não é obrigatoriamente a mesma numeração encontrada para sequências da mesma proteína obtida de diferentes organismos ou diferentes experimentos.

A ferramenta desenvolvida apresenta na sua saída as sequências alinhadas e renumeradas, de modo correto e consequentemente apresente a exibição das informações de maneira objetiva, através da representação visual acrescentado com o esquema de cor e a numeração correta para cada aminoácido, fazendo com que o usuário localize com mais clareza o aminoácido de interesse em um processo de alinhamento de sequências biológicas. Essas sequências poderão vir de diferentes experimentos e de diversas espécies.

A partir da ferramenta construída podemos destacar diversas vantagens da utilização da ferramenta comparada com as demais ferramentas existentes, dentre elas, é possível destacar a apresentação da régua horizontal enumerando cada aminoácido da sequência com a sequência que foi determinada como referência, após o processo de alinhamento. A régua horizontal é dinâmica e se adéqua de acordo com a sequência que é definida como referência, ajudando o usuário da ferramenta a encontrar de maneira rápida e eficaz o aminoácido de interesse. Nessa mesma perspectiva, a ferramenta utiliza informações representadas por meios de *tooltip's*, essas informações consistem em posição real que seria a posição original que o aminoácido se encontra na sequência e posição de alinhamento que se refere sobre a posição que o aminoácido se encontra logo após o processo

de alinhamento.

Por fim, é importante mencionar que a comparação de sequências proteicas é uma ferramenta essencial na procura da existência de relações de semelhança entre todo ou parte dessa sequência. Isso é muito comum quando temos uma sequência desconhecida e queremos identificar associando-a um grupo de proteínas de funções conhecidas, comparando essa sequência com outras de um banco de dados, também servem para a predizer as estruturas secundárias de proteínas ou para outras técnicas computacionais, como o docking e dinâmica molecular.

Palavras-chave: Sequência Biológica, Alinhamento Global, Alinhamento Local, Alinhamento Múltiplo, Essex.

ABSTRACT

JUNIOR, Wolmer Dias Quaresma. **Alignment of Biological Sequences - A new approach to optimal local alignments**. 2019. 97 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

The purpose of this dissertation is to propose a tool to deal with an important problem of computational biology, which corresponds to the location of a certain amino acid in a biological sequence so that it can obtain some biological meaning. Obtaining this information is often a complex process because the numbering of this amino acid in the sequence where the experiment was performed is not necessarily the same numbering found for sequences of the same protein obtained from different organisms or different experiments.

The developed tool presents in its output the correctly aligned and renumbered sequences and consequently displays the information in an objective way, through the visual representation added with the color scheme and the correct numbering for each amino acid, causing the user to locate with more clarity the amino acid of interest in a process of biological sequence alignment. These sequences may come from different experiments and from several species.

From the built tool we can highlight several advantages of using the tool compared to the other existing tools, among them it is possible to highlight the presentation of the horizontal ruler by enumerating each amino acid of the sequence with the sequence that was determined as reference after the alignment process . The horizontal ruler is dynamic and conforms according to the sequence that is defined as reference, helping the tool user to quickly and effectively find the amino acid of interest. In this same perspective, the tool employs information represented by tooltip's means, this information consists of the actual position that would be the original position that the amino acid is in the sequence and position of alignment that refers to the position that the amino acid is just after the alignment process.

Finally, it is important to mention that the comparison of protein sequences is an essential tool in the search for the existence of relations of similarity between all or part of that sequence. This is very common when we have an unknown sequence and we want

to identify it by associating it with a group of proteins of known functions, comparing this sequence with others of a database, also serve to predict the secondary structures of proteins or to other computational techniques like the docking and molecular dynamics.

Keywords: Biological Sequence, Global Alignment, Local Alignment , Multiple Alignment, Essex.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PDB - *Protein Data Bank* - Banco de dados de proteínas.

NW - Needleman-Wunsch - Algoritmo que realiza o alinhamento Global.

SW - Smith-Waterman - Algoritmo que realiza o alinhamento Local.

FASTA - Formato em texto para representar sequências.

UFMG - Universidade Federal de Minas Gerais.

VERMONT - ViewER MutatiON Tool - Ferramenta de exibição de mutações.

PROMALS3D - Ferramenta para alinhamento de sequências.

AA - Aminoácidos.

REF - Referência.

SEQ - Sequência.

ESSEX - Ferramenta Essex para alinhamento de Sequências.

EMBOSS - Ferramenta para alinhamento Global e Local.

HTML - *HyperText Markup Language* - Linguagem utilizada para criação de sites.

CSS - *Cascading Style Sheets* - Linguagem que descreve o estilo de uma página de HTML.

ID - Identidade

GAP - Descreve lacuna, vão ou fenda.

API - *Application Programming Interface* - Interface de Programação de Aplicativos.

TXT - Extensão de Arquivo de Texto.

BLOSUM - *BLOcks of Amino Acid SUbstitution Matrix* - Matriz de substituição usada para o alinhamento de sequências.

BLAST - *Basic Local Alignment Search Tool* - Algoritmo para comparar informações de sequências biológicas

CÓD - código.

SUMÁRIO

1 INTRODUÇÃO

A bioinformática é definida como uma área científica interdisciplinar que desenvolve métodos para armazenamento, recuperação, organização e análise de dados biológicos. Seu principal objetivo é desenvolver ferramentas de software para produzir informações relevantes a partir destes dados (BALDI; BRUNAK, 2001).

Nesse cenário, uma das funções mais essenciais em bioinformática é realizar a comparação da mesma espécie ou espécie distintas. Um modo para realizar essa comparação é através da aplicação de alinhamentos de sequências biológicas (CARLOS; MEIDANIS, 1997). Esses alinhamentos são utilizadas em várias áreas da bioinformática, dentre elas é possível citar o alinhamento de novos dados de sequenciamento e estudos de estruturas de proteínas.

Com uma frequência cada vez maior, são realizados estudos onde dados de sequenciamento de uma determinada espécie e origem experimental são alinhados com dados de outra espécie obtidos talvez com uma outra técnica experimental. Além disso, sequências de aminoácidos de uma determinada proteína para a qual se tem disponível sua estrutura tridimensional, também podem fazer parte de tal alinhamento. Este tipo de estudo, onde a área da bioinformática que estuda as estruturas tridimensionais da proteína e a genômica se encontram é cada vez mais comum, pois a quantidade de dados de sequenciamento disponíveis é consideravelmente superior a quantidade de dados estruturais. Desse forma, é natural a utilização de proteínas similares para estudar o efeito de mutações em sequência, ou para estudar interações proteína ligante, onde a proteína em estudo não possui a estrutura tridimensional definida.

Apesar de termos disponíveis diversas ferramentas que realizam o alinhamento de

sequências biológicas, nas nossas pesquisas não encontramos a disposição uma ferramenta que possibilita a renumeração do elementos da sequência. A renumeração poderia facilitar em muito a rápida localização de um elemento de uma sequência de referência em outras sequências alinhadas com a mesma.

1.1 Motivação

No entendimento sobre funções biológicas é essencial ter o conhecimento no processo de comparar diversas sequências biológicas. Tendo como exemplo, ter interesse em saber qual o ancestral de uma determinada sequência ou ver a similaridade das sequências para estudos de filogenias, pois obtendo essa informação talvez descobriríamos a similaridade entre espécies diferentes e fundamentando algum sentido biológico (PROSDOCIMI et al., 2002). Também é possível alinhar sequências de pacientes para comparar com sequências genômicas a fim de auxiliar em diversos tratamentos de doenças, descobrindo a eficácia de uma droga para determinados tratamentos, podendo assim serem indicadas, substituídas ou não recomendadas (BRITO, 2003).

Além das aplicações tradicionais de um resultado de alinhamento, a identificação da posição correta de um aminoácido específico de uma dada sequência proteíca em relação a outras sequências de origens experimentais diversas ou então de diferentes organismos é muito importante em um processo de alinhamento de sequências biológicas (FARIA, 2013).

Outro fator relevante é que a partir de um alinhamento podemos concluir se duas ou mais sequências estão evolutivamente associadas ou não. Além disso, os alinhamentos também servem para auxiliar na predição de estruturas secundárias e terciárias de proteína baseadas em homologia (TELLES; ALMEIDA; MARTINEZ, 2005).

De acordo com MCREE (1999), sequências de origens diferentes podem apresentar problemas relacionados aos experimentos utilizados para gerar estruturas tridimensionais, tendo como exemplo a cristalografia, originando falhas e consequentemente essas falhas poderão gerar problemas na contagem de aminoácidos. Algumas técnicas computacionais, como docagem molecular e dinâmica molecular, geralmente indicam aminoácidos

de interesse a serem investigados em mais detalhes. Tradicionalmente, alinhamento de sequências são utilizados em bioinformática para descobrir sequências homólogas para que seja possível predizer uma informação funcional, estrutural e evolucionária das sequências biológicas da proteína correspondente. Isso é muito comum quando temos uma sequência desconhecida e queremos identificar associando-a um grupo de proteínas de funções conhecidas, comparando essa sequência com outras de um banco de dados (ROCHA, 2011).

Porém, a numeração deste aminoácido na sequência onde o experimento foi realizado não é necessariamente a mesma numeração encontrada para sequências da mesma proteína/gene obtida de diferentes organismos e ou diferentes experimentos. Essas diferenças possivelmente ocorrem devido a serem sequências de espécies diferentes, por vir de diversas fontes e lugares e também sequências que apresentaram alguns problemas experimentais. Esta diferença de numeração dificulta muito a localização de aminoácidos em sequências de diferentes origens. Frequentemente, biólogos precisam fazer esse procedimento manualmente, no intuito de localizar um determinado aminoácido em uma sequência, para que o mesmo possua algum significado biológico. No entanto, é necessário obter uma identificação rápida e correta de regiões de interesse e mesmo de maneira manual é possível que não esteja correta.

Na atualidade, existem diversas ferramentas para realizar os alinhamentos de origem biológica. Entretanto, nossa proposta viabiliza encontrar a posição de um aminoácido de interesse utilizando um processo de alinhamento de sequências biológicas. Essas sequências por sua vez, podem ser de organismo ou de experimentos diferentes e de diversas fontes (FARIA, 2013).

Para resolver esse problema, elaboramos uma ferramenta que permite ao usuário inserir sequências de origens diferentes e locais distintos e posteriormente mostra a renumeração automática de todos os aminoácidos das sequências inseridas no processo de alinhamento, de maneira que o usuário da ferramenta localize de maneira precisa e rápida o aminoácido pretendido.

1.2 Objetivos

Desenvolver uma ferramenta para realizar o alinhamento e renumeração automática de cada aminoácido correspondente na sequência biológica de acordo com a sequência de referência definida pelo usuário.

1.3 Objetivos Específicos

- Atenuar o esforço humano para solucionar um problema importante da bioinformática, o alinhamento de sequências de diferentes naturezas.
- Permitir que o usuário possa incluir várias sequências de diferentes bases de dados e formatos.
- Possibilitar que o usuário, através da ferramenta, possa realizar os alinhamentos global, local e múltiplo das sequências inseridas.

1.4 Organização do Texto

Este trabalho está organizado da seguinte forma:

- O Capítulo 2 refere a fundamentação teórica, correlacionando estudos sobre aminoácidos, proteínas e suas respectivas estruturas. Além disso, é apresentada uma abordagem sobre a diferença entre as sequências da proteína extraída do arquivo PDB e de seu arquivo padrão Fasta. Ainda nesse capítulo, é feita uma contextualização sobre biologia evolutiva e abordaremos posteriormente sobre a definição de alinhamento e também os três modos de executar os alinhamentos: Alinhamento Global, Alinhamento Local e Alinhamento Múltiplo de Sequências e os possíveis problemas que possam ocorrer durante esse processo. Por fim, descreveremos sobre Algoritmos de Programação Dinâmica.
- O Capítulo 3, por sua vez, lista trabalhos relacionados como Vermont, Emboss Stretcher/Matcher, ClustalW2, Promals3D, Clustal Omega e ICM Browser Pro e a apresenta uma tabela comparativa entre eles com o trabalho proposto.

- A metodologia proposta, as ferramentas utilizadas e a ferramenta proposta neste trabalho são descrita no Capítulo 4.
- No Capítulo 5, apresenta os estudos de casos realizado.
- No Capítulo 6, mostra as considerações finais contendo a discussão e a conclusão.

Ao final, no Capítulo 7, elaboramos um tutorial para o usuário utilizar o Essex.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Aminoácidos

A estrutura geral dos aminoácidos envolvem, pelo menos, um grupo amina H_2N e um grupo carboxílico $-COOH$, um átomo de hidrogênio H, um carbono alfa C_α e uma cadeia lateral, que é representada pela letra R (HRUBY, 1986). Esses aminoácidos são encontradas nos organismos vivos na forma de peptídeos (dois ou mais aminoácidos) e proteínas (acima de 51 aminoácidos). Nesse contexto, a cadeia lateral é encarregada por determinar a identidade dos aminoácidos existentes (ACID, 2004). A figura 1 mostra a estrutura química básica dos 20 aminoáciados.

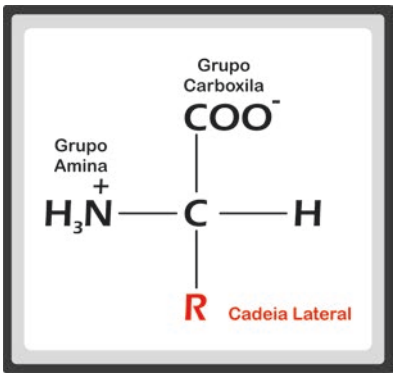

Figura 1: Estrutura Básica do Aminoácido.

Fonte: Adaptado de (ACID, 2004)

A nomenclatura dos aminoácidos é formada por abreviações de uma ou três letras. Na tabela 1, temos a lista dos 20 aminoácidos existentes nas proteínas.

Nome	Abreviatura (3 letras)	Abreviatura (1 letra)
Alanina	Ala	A
Cisteína	Cys	C
Asparato	Asp	D
Glutamato	Glu	E
Fenilalanina	Phe	F
Glicina	Gly	G
Histidina	His	H
Isoleucina	Ile	I
Lisina	Lys	K
Leucina	Leu	L
Metionina	Met	M
Asparagina	Asn	N
Prolina	Pro	P
Glutamina	Gln	Q
Arginina	Arg	R
Serina	Ser	S
Treonina	Thr	T
Valina	Val	V
Triptofano	Trp	W
Tirosina	Tyr	Y

Tabela 1: Lista de Aminoácidos e suas abreviaturas

2.2 Proteínas

As proteínas são polímeros sintetizados pelas células a partir de aminoácidos e constituem o principal produto direto da informação genética a partir da tradução do RNA mensageiro (VERLI, 2014). Elas são as mais importantes das macromoléculas biológicas, são as moléculas mais abundantes da natureza. Estão presente em todo ser vivo e tem as mais variadas funções (FERRIER; HARVEY, 2011).

2.2.1 Estrutura da Proteína

Apesar de uma proteína ser uma cadeia linear de aminoácidos unidos por ligações covalentes, a sua forma e função são determinados pela sua estrutura tridimensional. Esse formato é determinado pela extensa associação de interações fracas individu-

ais formadas entre aminoácidos que não são necessariamente adjacentes na sequência primária (SCHULZ; SCHIRMER, 2013). São quatro níveis de estruturas para as proteínas: primárias, secundárias, terciárias e quaternárias e a sua complexidade vai ampliando de acordo com a estrutura que estamos trabalhando.

A estrutura primária corresponde ao primeiro nível de organização estrutural que a proteína possui. Refere-se a cadeia principal da proteína construída pela ligação dos aminoácidos. A união de dois aminoácidos é feita através de uma ligação química chamada de peptídica. Por sua vez, a ligação polipeptídica remete a união de vários aminoácidos.

O próximo nível da estrutura proteica, chamado de estrutura secundária, é definido à medida que o comprimento das cadeias vai aumentando gradativamente e também por essa cadeia polipeptídica ser flexível (BRANDEN et al., 1999). As duas principais conformações da estrutura secundária da proteína são Alfa-hélice e Folha-beta conforme é mostrado na figura 2. Na alfa-hélice, a estrutura polipeptídica se dobra através da interação das ligações de hidrogênio gerando uma forma de espiral ou de uma hélice. Já na Folha-beta, é quando dois ou mais elementos de uma cadeia polipeptídica se alinham próximos aos outros, formando uma estrutura semelhante como uma folha dobrada que é mantida pelas ligações de hidrogênio (KABSCH; SANDER, 1983).

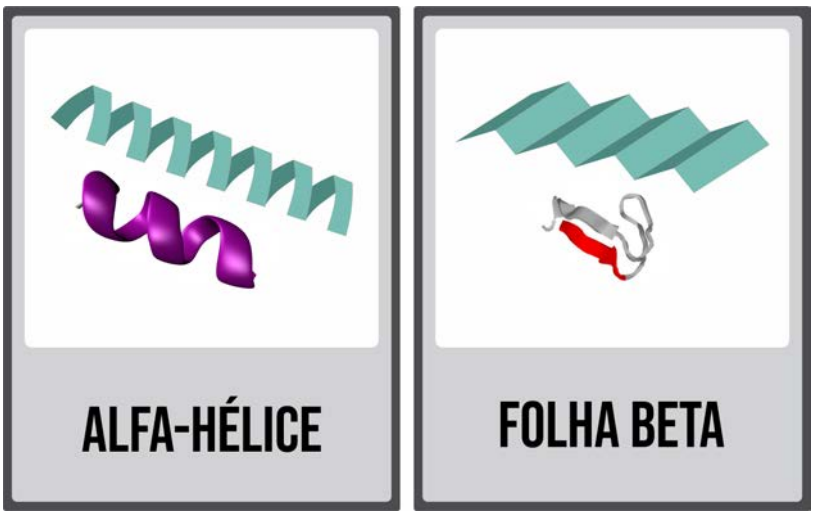

Figura 2: Alfa-hélice x Folha-beta.

Fonte: Dados do Autor.

A estrutura terciária é a forma como o dobramento da estrutura secundária se organiza no espaço de forma tridimensional. Também é estabilizada por ligações de hidrogênio e dissulfeto, o que garante maior estabilidade à proteína.

Muitas proteínas são formadas por mais de uma cadeia polipeptídica. A estrutura quaternária é a ligação de inúmeras estruturas terciárias que obtém formas espaciais bem definidas e se ajustam para formar a estrutura por inteiro de uma proteína (LESK, 2001). Na figura 3 exibimos exemplos de níveis das estruturas de uma proteína.

Figura 3: Níveis de Estruturas de Proteínas.

Fonte: Dados do Autor.

2.2.2 Sequência no formato Fasta

O arquivo FASTA contém a sequência de cada cadeia de aminoácidos ou nucleótidos padrão ou alterados, oferecendo detalhes básicos da sequência de uma proteína específica. Na primeira linha, precedido do símbolo >, há informações sobre a proteína ou gene, o ID, cadeias e outros detalhes sobre a finalidade da proteína. Sucessivamente, temos a sequência primária da proteína (BERMAN; NAKAMURA; HENRICK, 2010). A figura 4 mostra uma representação de uma sequência da proteína 1IWG no formato fasta.

```
>1IWG:A|PDBID|CHAIN|SEQUENCE
MPNFFIDRPIFAWVIAIIIMLAGGLAILKLPVAQYPTIAPPAVTISASYPGADAKTVQDTVTQVIEQNMNGIDNLMYMSS
NSDSTGTVQITLTFESGTDADIAQVQVQNKLQLAMPLLPQEVQQQGVSVEKSSSSFLMVVGVINTDGTMTQEDISDYVAA
NMKDAISRTSGVGDVQLFGSQYAMRIWMNPNELNKFQLTPVDVITAIKAQNAQVAAGQLGGTPPVKGQQLNASIIAQTRL
TSTEEFGKILLKVNQDGSRVLLRDVAKIELGGENYDIIAEFNGQPASGLGIKLATGANALDTAAAIRAELAKMEPFFPSG
LKIVYPYDTTPFVKISIHEVVKTLVEAIILVFLVMYLFLQNFRATLIPTIAVPVVLLGTFAVLAAFGFSINTLTMFGMVL
AIGLLVDDAIVVVENVERVMAEEGLPPKEATRKSMGQIQGALVGIAMVLSAVFVPMAFFGGSTGAIYRQFSITIVSAMAL
SVLVALILTPALCATMLKPIAKGDHGEGKKGFFGWFNRMFEKSTHHYTDSVGGILRSTGRYLVLYLIIVVGMAYLFVRLP
SSFLPDEDQGVFMTMVQLPAGATQERTQKVLNEVTHYYLTKEKNNVESVFAVNGFGFAGRGQNTGIAFVSLKDWADRPGE
ENKVEAITMRATRAFSQIKDAMVFAFNLPAIVELGTATGFDFELIDQAGLGHEKLTQARNQLLAEAAKHPDMLTSVRPNG
LEDTPQFKIDIDQEKAQALGVSINDINTTLGAAWGGSYVNDFIDRGRVKKVYVMSEAKYRMLPDDIGDWYVRAADGQMVP
FSAFSSSRWEYGSPRLERYNGLPSMEILGQAAPGKSTGEAMELMEQLASKLPTGVGYDWTGMSYQERLSGNQAPSLYAIS
LIVVFLCLAALYESWSIPFSVMLVVPLGVIGALLAATFRGLTNDVYFQVGLLTTIGLSAKNAILIVEFAKDLMDKEGKGL
IEATLDAVRMRLRPILMTSLAFILGVMPLVISTGAGSGAQNAVGTGVMGGMVTATVLAIFFVPVFFVVVRRRFSRKNEDI
EHSHTVDHHHHHH
```

Figura 4: Sequência FASTA da proteína código 1IWG

Fonte: Dados do Autor.

2.2.3 Arquivo no formato PDB

Os arquivos PDB contêm informações de sequências e coordenadas atômicas, esses arquivos são armazenados em bancos de dados de proteínas e podem ser lidos por diversos programas. A descrição completa do arquivo PDB oferece uma riqueza de informações, incluindo autores, referências bibliográficas e o método de determinação da estrutura. Essas informações são representadas por linhas em um formato de arquivo de texto (BERMAN et al., 2006). A próxima seção mostrará algumas informações contidas em um arquivo PDB, composto por vários tipos de registros, organizados em uma ordem específica para descrever uma estrutura

2.2.3.1 Title Section

Esta seção contém registros usados para explicar o experimento e as macromoléculas biológicas presentes.

- HEADER: O registro *HEADER* identifica exclusivamente uma entrada do PDB por meio do campo idCode. Por fim, contém a data em que as coordenadas foram depositadas no arquivo do PDB.

- TITLE: Trata do título que é dado para o experimento ou análise representado.

- SOURCE: O registro *SOURCE* especifica a fonte biológica e/ou química de cada

molécula biológica. Alguns casos em que a entrada contém um medicamento ou inibidor autônomo, a informação de origem desta molécula aparecerá neste registro. As fontes são descritas tanto pelo nome comum quanto pelo nome científico, por exemplo, gênero e espécie.

- KEYWDS: Remete a *Keywords* que são um conjunto de termos relevantes para a entrada

- REVDAT: Os registros REVDAT contêm um histórico das modificações feitas desde o seu lançamento.

- EXPDTA: O registro EXPDTA apresenta informações sobre o experimento, identificando a técnica experimental usada. Pode se referir ao tipo de radiação e amostra.

- AUTHOR: Esse registro informa os nomes das pessoas responsáveis pelo conteúdo.

- JRNL: Contém a citação da literatura principal que descreve o experimento que resultou no conjunto de coordenadas depositadas.

- REMARK: Registros *REMARK* apresentam detalhes experimentais, anotações, comentários e informações não incluídos em outros registros.

2.2.3.2 Primary Structure Section

Retrata a estrutura primária de um arquivo no formato PDB, contém a sequência de resíduos em cada cadeia da(s) macromolécula(s). Esses registros incorporados são identificadores de cadeia e números de sequência que permitem que outros registros sejam vinculados à sequência (BERMAN et al., 2009).

- DBREF: O DBREF fornece links de referência cruzada entre as sequências do PDB (o que aparece no registro SEQRES) e uma sequência de banco de dados correspondente.

- SEQADV: Esse registro identifica as diferenças entre as informações da seqüência nos registros SEQRES da entrada do PDB e a entrada do banco de dados de

sequência fornecida no DBREF. Nenhuma suposição é feita a respeito de qual banco de dados contém os dados corretos.

- SEQRES: Os registos SEQRES contêm uma listagem dos componentes químicos consecutivos ligados de forma covalente de uma forma linear para formar um polímero. Os componentes químicos incluídos nesta listagem podem ser aminoácidos padrão ou modificados e resíduos de ácido nucleico. Todo o arquivo com formato PDB existe a coluna "SEQRES" no qual obtemos uma lista que remete a sequência primária das moléculas. Esta informação também está disponível como um download no formato FASTA. (BRANDT; HERINGA; LEUNISSEN, 2008).

2.2.3.3 Secondary Structure

Apresenta a estrutura secundária da proteína descrita em um arquivo no formato PDB. Descreve hélices e folhas encontradas em estruturas de proteína e polipeptídeo (BERMAN et al., 2007).

- HELIX: Os registros *HELIX* são usados para identificar a posição das hélices na molécula. As hélices são nomeadas, numeradas e classificadas por tipo.

- SHEET: os *SHEET* são usados para identificar a posição das folhas na molécula. As folhas, assim como as hélices, são nomeadas e numeradas.

2.2.3.4 Coordinate Section

Esta seção de coordenadas contém o conjunto de coordenadas atômicas, descrevendo a estrutura terciária da proteína apresentada no arquivo PDB (BERMAN et al., 2009).

- ATOM: O *ATOM* contém uma lista das coordenadas (x, y e z) de todos os átomos da estrutura 3D de uma proteína e núcleotideos. O arquivo PDB por sua vez, é muito maior do que um arquivo FASTA, pois precisa abranger informações sobre cada átomo na proteína, além da massa e posição dos átomos no espaço tridimensional. (BERMAN et al., 2006). A figura 5 mostra como é representado esses registros em um arquivo PDB.

```
ATOM      1  N   ASP A   7      56.989  30.960 269.629  1.00127.37
ATOM      2  CA  ASP A   7      58.446  30.827 269.347  1.00127.37
ATOM      3  C   ASP A   7      58.929  29.425 269.711  1.00127.37
ATOM      4  O   ASP A   7      58.730  28.475 268.954  1.00127.37
ATOM      5  CB  ASP A   7      59.234  31.869 270.147  1.00136.61
ATOM      6  CG  ASP A   7      58.748  33.285 269.900  1.00136.61
ATOM      7  OD1 ASP A   7      59.333  34.225 270.478  1.00136.61
ATOM      8  OD2 ASP A   7      57.779  33.458 269.130  1.00136.61
ATOM      9  N   ARG A   8      59.561  29.305 270.874  1.00206.80
ATOM     10  CA  ARG A   8      60.070  28.023 271.346  1.00206.80
ATOM     11  C   ARG A   8      59.415  27.642 272.672  1.00206.80
ATOM     12  O   ARG A   8      59.618  28.307 273.688  1.00206.80
ATOM     13  CB  ARG A   8      61.590  28.088 271.516  1.00185.81
ATOM     14  CG  ARG A   8      62.343  28.384 270.228  1.00185.81
ATOM     15  CD  ARG A   8      63.849  28.369 270.442  1.00185.81
ATOM     16  NE  ARG A   8      64.284  29.396 271.383  1.00185.81
ATOM     17  CZ  ARG A   8      65.551  29.607 271.725  1.00185.81
```

Figura 5: Exemplo de Coluna ATOM da proteína código PDB: 1IWG
Fonte: Dados do Autor.

Na figura 5, é possível observar que cada linha começa com o tipo de registro ATOM, o número de série do átomo é o próximo item da coluna. O nome do átomo é localizada na terceira coluna e é representada com um ou dois caracteres que consistem no símbolo químico. Todos os nomes que começam com C são átomos de carbono, N indica um nitrogênio e O indica oxigênio. No segundo caractere da terceira coluna, fica situado os resíduos de aminoácidos e mostra o código indicador de afastamento, que é transliterado de acordo com a tabela 2:

Símbolo	**Letra**
α	A
β	B
γ	G
δ	D
ε	E
ζ	Z
η	H

Tabela 2: Lista dos códigos indicadores e suas representações

Na coluna 4 identificamos qual aminoácido da proteína. Neste exemplo, o primeiro aminoácido na cadeia é ASP (Asparato) e o segundo aminoácido da cadeia é ARG (Arginina). A próxima coluna, contém o identificador da cadeia, neste caso A.

Na coluna 6 contém o número de sequência de resíduos. Observe que à medida que o resíduo muda de Asparato para Arginina, o número de resíduos muda de 7 para 8. Dois aminoácidos semelhantes podem estar próximos um ao outro, portanto, o número de resíduos é importante para distinguirmos um ao outro.

Os próximos três campos de dados mostram respectivamente os valores das coordenadas X, Y e Z em que os átomos estão situados. As últimas colunas, que pode variar de dois ou três campos retratam o fator de temperatura.

- HETATM: Não-polímero ou outras coordenadas químicas "não-padrão", como moléculas de água ou átomos apresentados nos grupos HET, usam o tipo de registro *HETATM*. Eles também apresentam o fator de ocupação e temperatura para cada átomo (BERMAN et al., 2009).

- TER: O registro TER indica o final de uma lista de registros ATOM/HETATM para cada cadeia.

2.3 Biologia Evolutiva

Alinhamentos de sequências biológicas tendem a indicar uma homologia entre as sequências ou, pelo menos, mostrar a similaridade e sua significância. Os termos homologia, semelhança e identidade têm significados diferentes no tema da análise de sequências biológicas.

- Homologia: descreve uma relação evolutiva entre duas sequências que poderão corresponder a proteínas de funções homologas em diferentes organismos que partilham um ancestral comum (FITCH, 2000).

  Os genes homólogos possuem três eventos distintos:

  Ortólogos são aqueles genes que divergiram após um evento de especiação. Em outras palavras, genes ortólogos são "genes iguais" em diferentes organismos, possuem uma origem homóloga e funções similares (SONNHAMMER; KOONIN, 2002)

Por outro lado, dois genes homólogos são identificados parálogos quando foram gerados por um evento de duplicação dentro do genoma de uma mesma espécie, ocupando diferentes posições no mesmo genoma. Portanto, têm funções distintas.

Em relação aos genes xenólogos, estes originam-se quando os genes homólogos são o produto da transferência horizontal de genes entre duas espécies e, dependendo do novo ambiente para o qual o gene foi movido, podem ter funções semelhantes ou distintas (FITCH, 1970).

- Semelhança: Descreve o grau de relação entre duas sequências, que não depende do seu contexto ou significado biológico, utilizando um método matemático que considera a probabilidade do alinhamento ter ocorrido por acaso.

- Identidade: é expressado normalmente em um percentual de identidade entre duas sequências, sendo compreendida entre o número total de resíduos idênticos e o número total de resíduos do alinhamento.

2.4 Definição de Alinhamento de Sequências

O alinhamento de sequências é estabelecido pela comparação de duas ou mais sequências, com propósito de descobrir uma série de caracteres individuais ou grupo de caracteres que estejam na mesma ordem nas duas sequências. A similaridade é muito importante para identificar ou quantificar o quão similar é uma determinada sequência à outra (TICONA, 2003).

Esse processo de alinhar duas ou mais sequências biológicas é embasada na técnica de algoritmo de programação dinâmica, que busca o melhor alinhamento entre as sequências. A técnica consiste na construção de uma matriz de comparação de sequências a serem alinhadas, atribuindo um *score* (valor de pontuação).

O valor de pontuação é computado com o objetivo de penalizar as diferenças entre os aminoácidos nas sequências inseridas para a realização do alinhamento e privilegiar as similaridades. Alguns casos, contam com sequências de tamanhos diferentes. Logo, o algoritmo de programação dinâmica insere *gaps* (espaços) nas sequências com o intuito

de que as sequências estejam com o mesmo tamanho após o processo de alinhamento.

A figura 6 apresenta o resultado de um possível processo de alinhamento, analisando as sequências seguintes: TKVNGSLETA e TKVGTLETA.

Figura 6: Exemplos de sequências.

Fonte: Dados do Autor.

É possível observar que as sequências são bem semelhantes. As diferenças são: um aminoácido (N) a mais na primeira sequência. Desta forma, na segunda sequência ocorreu a necessidade da inserção de um *gap* para que mais aminoácidos coincidam e consequentemente obter um resultado melhor no alinhamento realizado. Uma nova diferença é uma troca de (S) por (T) na quinta posição da direita para a esquerda (ZAFALON, 2009).

Vale ressaltar que para qualquer realização de alinhamento de sequências utilizamos a matriz de substituição. No momento em que acontece a coincidência, similaridade, entre os aminoácidos nas mesmas posições nas diferentes sequências, chamamos de *match*. Ao contrário, quando não existe a similaridade ou quando existe a inserção de *gap* naquela mesma posição, ocorre um *mismatch*. Entretanto, o resultado de um alinhamento pode ser completamente diferente dependendo da matriz de substituição utilizada.

Durante o processo de alinhamento de sequências, as matrizes de substituição são especialmente usadas pois com elas podemos escolher qual características dos aminoácidos queremos levar em conta na comparação, para um determinado alinhamento pode ser importante manter as características de polaridade, por exemplo, existem aminoácidos com cargas polares, apolares ou sem carga, logo, a mudança de um aminoácido que apresenta uma determinada característica para outro da mesma característica é menos significante na função de uma determinada proteína do que mudanças de aminoácidos que apresentam

características diferentes. As matrizes de substituição mais conhecidas para sequências proteicas são matrizes BLOSUM (Blocks Substitution Matrix). Elas, por exemplo, são baseadas na observação das frequências de substituição em blocos de alinhamentos locais de proteínas relacionadas (PROSDOCIMI et al., 2002).

Logo, é possivel concluir que para alinhar duas sequências biológicas é necessário determinar quais as medidas que serão tomadas para calcular o *score* do alinhamento, o melhor alinhamento será aquele que obtiver maior *score*. Os métodos de alinhamentos existentes são: Global, Local e Múltiplo (CARLOS; MEIDANIS, 1997).

A figura 7 descreve duas sequências (seq1 e seq2) que são utilizadas para exemplificar os alinhamentos global e local nas seções 2.4.1 e 2.4.2.

Figura 7: Exemplos de sequências

Fonte: Dados do Autor.

2.4.1 Alinhamento Global

O alinhamento global significa que é feito um experimento de alinhar toda extensão da sequência, utilizando o máximo de caracteres possíveis. As sequências semelhantes e aproximadamente do mesmo comprimento são candidatos adequados para realizar o alinhamento global (GOLLERY, 2005). A figura 8 exibe o resultado do alinhamento global.

Figura 8: Exemplos de Alinhamento Global.

Fonte: Dados do Autor.

2.4.2 Alinhamento Local

No alinhamento local, procura-se alinhar fragmentos com as maiores semelhanças locais. Não é utilizado a sequência em toda sua extensão, o alinhamento para nas extremidades dessas regiões e sua descoberta é dado a partir da maior pontuação obtida do que a utilização de toda a extensão do alinhamento. É apto para sequências com fragmentos de similaridade descontínuas, que possuem divergência de tamanho ou que possuem uma região de caracteres bastante conservada (ALTSCHUL et al., 1990). A figura 9 mostra o resultado do alinhamento local.

Figura 9: Exemplos de Alinhamento Local.

Fonte: Dados do Autor.

2.4.3 Alinhamento Múltiplo de Sequências

Os alinhamento múltiplos de sequências refere-se ao procedimento de alinhar um grupo de sequências biológicas. A extensão dos algoritmos de programação dinâmica (NEEDLEMAN; WUNSCH, 1970) ao alinhamento múltiplo é computacionamente cara e inviável para alinhar um conjunto de sequências (LIPMAN; ALTSCHUL;

KECECIOGLU, 1989), por isso, muitos algoritmos foram desenvolvidos para realizar esse processo. A figura 10 descreve três sequências (seq1, seq2 e seq3) que são utilizadas para exemplificar o alinhamento múltiplo na figura 10.

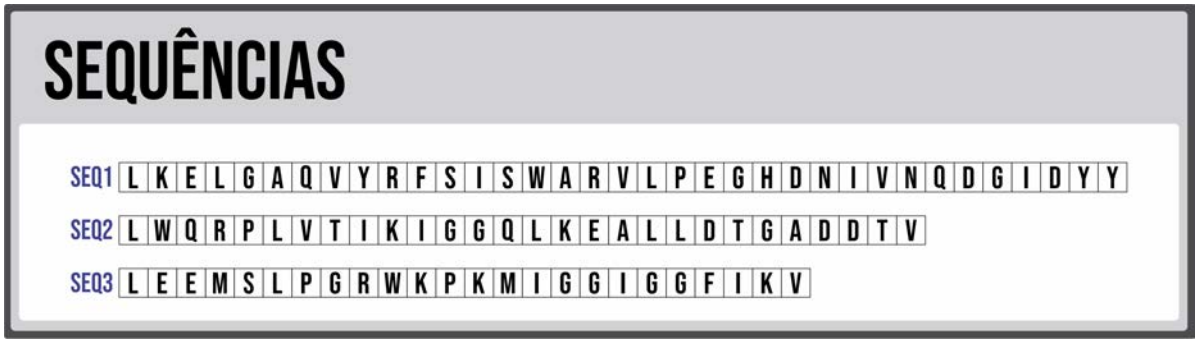

Figura 10: Exemplos de Alinhamento Múltiplo.

Fonte: Dados do Autor.

Consiste em um método progressivo que realiza uma série de alinhamentos em pares de sequências ou grupos já pré-alinhados. Sucessivamente, a ordem desses alinhamentos em pares é guiada por uma árvore, de modo que sequências similares tendem a ser alinhadas antes de sequências divergentes. Entretanto, essa metodologia não garante uma solução ideal ou uma melhor precisão do que alinhamentos globais realizado par-a-par (FENG; DOOLITTLE, 1987). A figura 11 exibe o resultado do exemplo de alinhamento múltiplo.

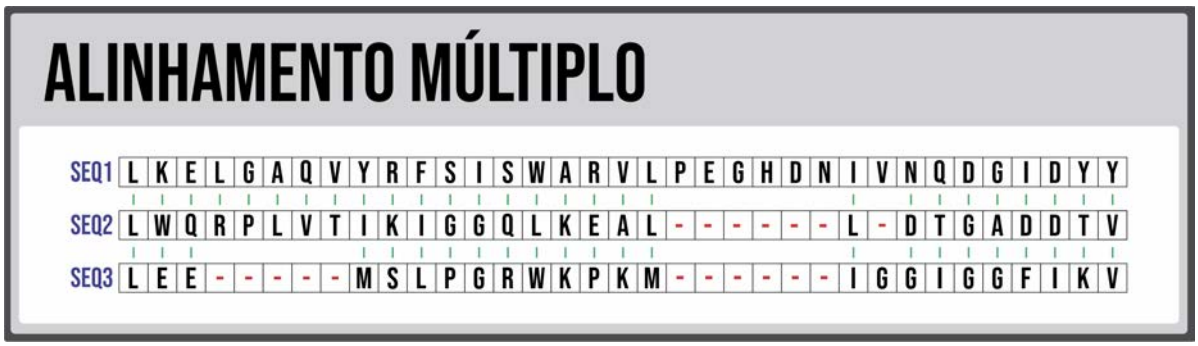

Figura 11: Exemplo de Resuldado de Alinhamento Múltiplo.

Fonte: Dados do Autor.

2.4.4 Possíveis Problemas no Alinhamento

Embora seja cada vez mais frequente a discussão sobre a relevância e eficiência dos métodos de alinhamentos de sequências biológicas, é possível perceber que há uma série de problemas que deverão ser enfrentados em cada alinhamento a ser realizado (SIPPL; WIEDERSTEIN, 2008).

Alguns desses problemas é devido as sequências que irão ser utilizadas para a realização do alinhamento que possuem tamanhos divergentes, também é costumeiro apresentar apenas uma pequena região nas sequências que é capaz de realizar o alinhamento e possivelmente essas regiões de tamanhos varáveis podem ter sofrido mutações, dentre elas podemos citar: substituições, inserções e remoções (BRITO, 2007).

2.5 Algoritmo de Programação Dinâmica

Com o objetivo de descobrir um alinhamento entre duas sequências, o procedimento mais utilizado é fundamentado no algoritmo de programação dinâmica. Tendo como objetivo procurar o melhor alinhamento entre duas *Strings* de caracteres que é embasada de uma matriz de comparação que representa as sequências de aminoácidos a serem alinhadas (BRITO, 2003).

Na tentativa de encontrar o melhor alinhamento, a técnica de programação dinâmica concede uma pontuação para cada par de aminoácidos alinhados, a pontuação é dada de maneira a penalizar as diferenças e beneficiar as similaridades. Sucessivamente, após construir essa matriz de comparação, é obtido o melhor alinhamento entre as sequências (BERGER; ROZENER, 1998).

2.5.1 Algoritmo de Needleman-Wunsch

Conforme NEEDLEMAN; WUNSCH (1970), foi apresentado um algoritmo baseado na programação dinâmica para o problema de alinhamento global onde é executado em dois estágios: cálculo da matriz e o $traceback$.

Conforme descrito na seção 2.4, um alinhamento é declarado global na ocasião em que serão utilizados todos os aminoácidos das sequências que serão alinhadas. Para localizar

os alinhamentos locais, observa os $scores$ realizado na tentativa de parear as sequências, esses aminoácidos no momento em que efetuam o alinhamento podem coincidir ou não.

Havendo a coincidência ela é chamada de *Match* e é avaliada positivamente e quando a coincidência não acontece ela é chamada de *Mismatch* e consecutivamente avaliada negativamente. Existem também os *Gap's*, que é no momento em que são penalizadas as inserções e deleções, permitindo retirar um significado biológico de acordo acontece esse tipo de ocorrência (NEEDLEMAN; WUNSCH, 1970).

2.5.2 Algoritmo Smith-Waterman

A principal diferença entre alinhamentos global e local ocorre quando o alinhamento local tem como seu objetivo alinhar fragmentos das sequências ao invés de alinhar elas por inteiros, esses fragmentos por sua vez, são regiões de alto grau de similaridade.

O algoritmo proposto por SMITH; WATERMAN (1981) é baseado no algoritmo de Needleman-Wunsch (NW), no entanto adaptado por meio de programação dinâmica para lidar com o problema de alinhamento local. No algoritmo NW, o alinhamento considerado ótimo pode possuir um score negativo, contando que seja o maior obtido. Para impossibilitar que isso aconteça, o algoritmo de Smith-Waterman teve sua equação alterada para que impeça valores menores que zero.

Caso isso aconteça, atribuí um valor zero caso o score seja negativo, logo, faz com que o algoritmo busque a maior região de similaridade entre duas sequências. Além disso, como SW é utilizado para alinhamentos locais, não é possível alinhamentos com grandes números de inserções, remoções e substituições (SMITH; WATERMAN, 1981).

2.5.3 Algoritmo ClustalW

O ClustalW é um algoritmo que realiza o alinhamento múltiplo de sequências biológicas que é embasado essencialmente em três passos: alinhamento par-a-par, construção de árvores guia e alinhamento progressivo. O algoritmo utilizado por este programa é heurístico e usa uma estratégia de alinhamento progressivo para a obtenção de um ótimo alinhamento múltiplo (LARKIN et al., 2007).

Com isso é possível a construção de árvores classificadoras, pegando as sequências

que são pretendidas no processo de alinhamento. A ordem com que as sequências são adicionadas ao alinhamento múltiplo é dedicida através da estratégia que o ClustalW utiliza, ou seja, todos os pares de sequências são comparados entre si agrupando-se as sequências mais similares que são as que apontam maior score no alinhamento simples (AIYAR, 2000).

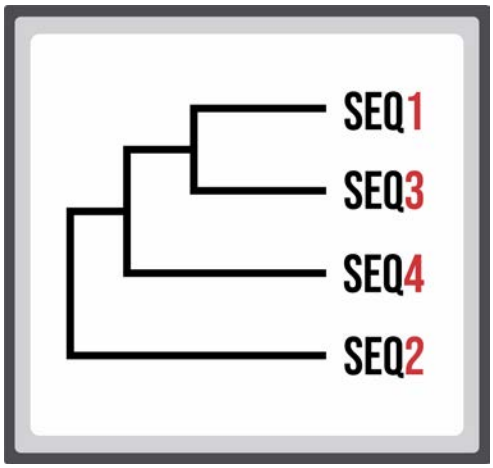

Figura 12: Representação do esquema da árvore guia.

Fonte: Dados do Autor.

Segundo a figura 12, o ClustalW realiza o alinhamento de sequências que têm maior score, especificamente Seq1 e Seq3 (Seq1, Seq3). Ao final desse processo de alinhamento, é comparado a sequência obtida com a sequência Seq4, sequência mais similar com a resultante do alinhamento anterior, e volta a alinhar (Seq1,Seq3),Seq4) (HIGGINS; THOMPSON; GIBSON, 1996).

As sequências são adicionadas ao alinhamento múltiplo respeitando à estrutura da árvore guia, das folhas para a raiz. A ilustração em árvore mostra assim a ordem pela quals as sequências são escolhidas (CHENNA et al., 2003).

3 TRABALHOS RELACIONADOS

3.1 Vermont (ViewER MutatiON Tool)

O Vermont é um *framework* de visualização interativa que permite aos usuários explorar um conjunto diversificado de informações sobre resíduos e mutações, possibilitando a identificação de importantes mutações e suas possíveis funções proteica. Além disso, a ferramenta contribui na investigação, seleção e combinação de um conjunto de dados baseados em sequência e estrutura para estimar o impacto de mutações, técnicas de interação também que podem ajudar na análise (SILVEIRA et al., 2014).

As mutações ocorrem de maneira natural devido a evolução, modificando a sequência de resíduos de uma proteína, o que pode alterar eventualmente a sua função. Por isso, uma das questões significativas é compreender como essas mutações em resíduos de proteínas podem afetar ou não uma determinada função da proteína. Portanto, é necessário o uso de instruções computacionais confiáveis para auxiliar o entendimento das mutações e seus impactos (FASSIO et al., 2017).

O seu funcionamento é dado a partir de uma entrada indicada pelo usuário e um limite de similaridade pretendido ou o usuário insere uma lista de entradas PDB e qual método de alinhamento local será realizado, e a ferramenta pesquisa no Protein Data Bank[1] para averiguar estruturas similares. O Vermont realiza o processo de alinhamento e avisa ao usuário quando a tarefa for concluída através do e-mail que o usuário introduziu.

Na figura 13 observamos a tela inicial do *framework* Vermont.

[1] https://www.rcsb.org/

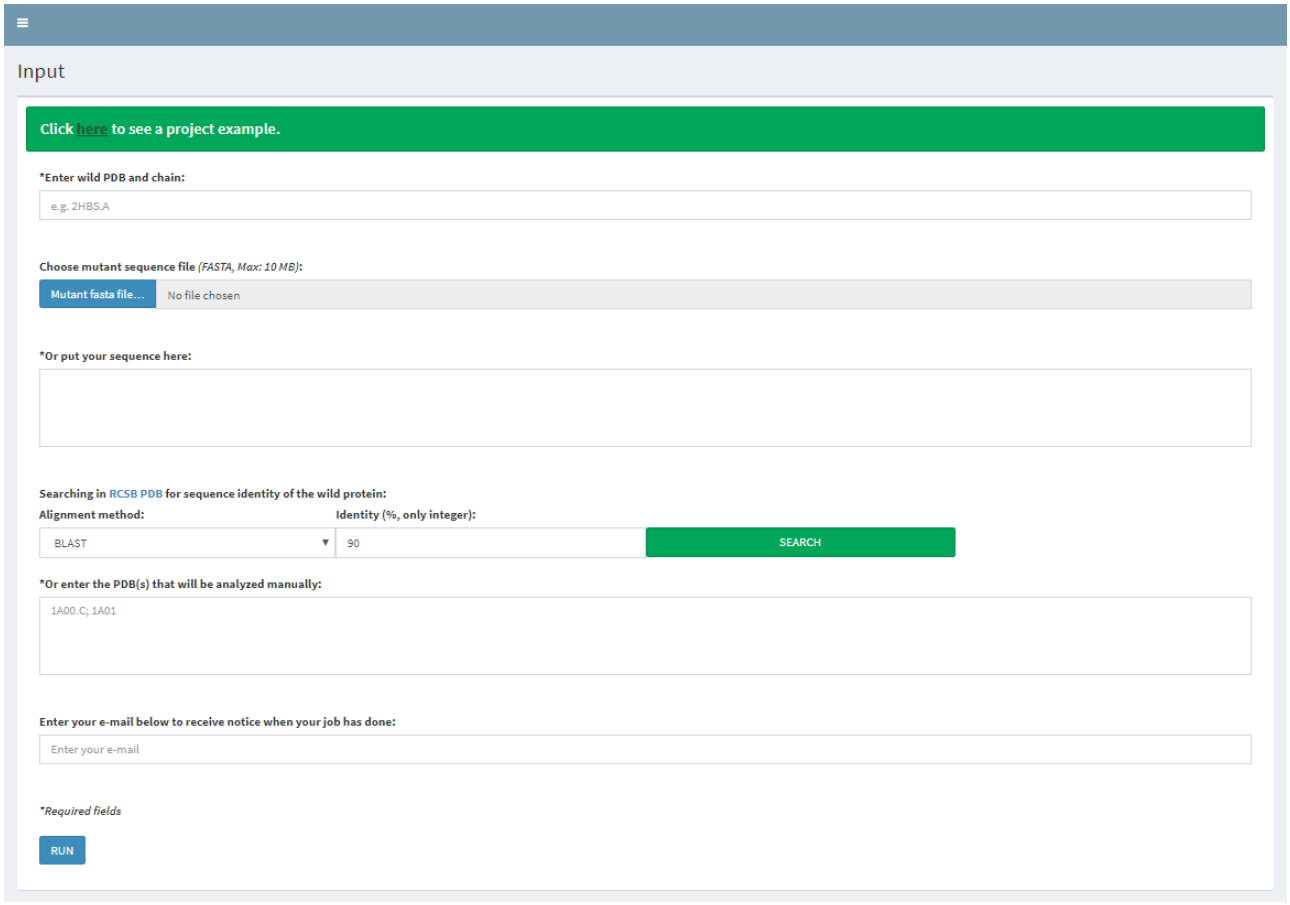

Figura 13: Tela inicial do VERMONT (Viewer Mutation Tool)

Fonte: Extraído de http://bioinfo.dcc.ufmg.br/vermont/results/view

Para os biólogos e especialistas da bioinformática, a representação visual é um fator positivo conforme é mostrado na figura 14, pois o Vermont utiliza a visualização de alinhamento de sequências múltiplas juntamente com gráficos de interação e visualizações estruturais moleculares de proteínas que foram adicionados para intensificar a análise de especialistas (SILVEIRA et al., 2014).

Somando-se a isso, os resíduos em geral possuem cores de acordo com um esquema de cores que é relacionado com as propriedades físico-químicas, observando a conservação em uma determinada coluna do alinhamento que foi realizado.

Figura 14: Tela de resultados do VERMONT (Viewer Mutation Tool)

Fonte: Extraído de http://bioinfo.dcc.ufmg.br/vermont/results/view

3.2 Emboss Stretcher (Global) / Matcher (Local)

Alinhamento de sequências biológicas é um processo que, dependendo da aplicação, pode levar um tempo considerável, variando com tamanho das sequências a serem alinhadas, além de utilizar muita memória (LI et al., 2015).

A figura 15 exibe a interface da ferramenta Emboss Stretcher/Matcher.

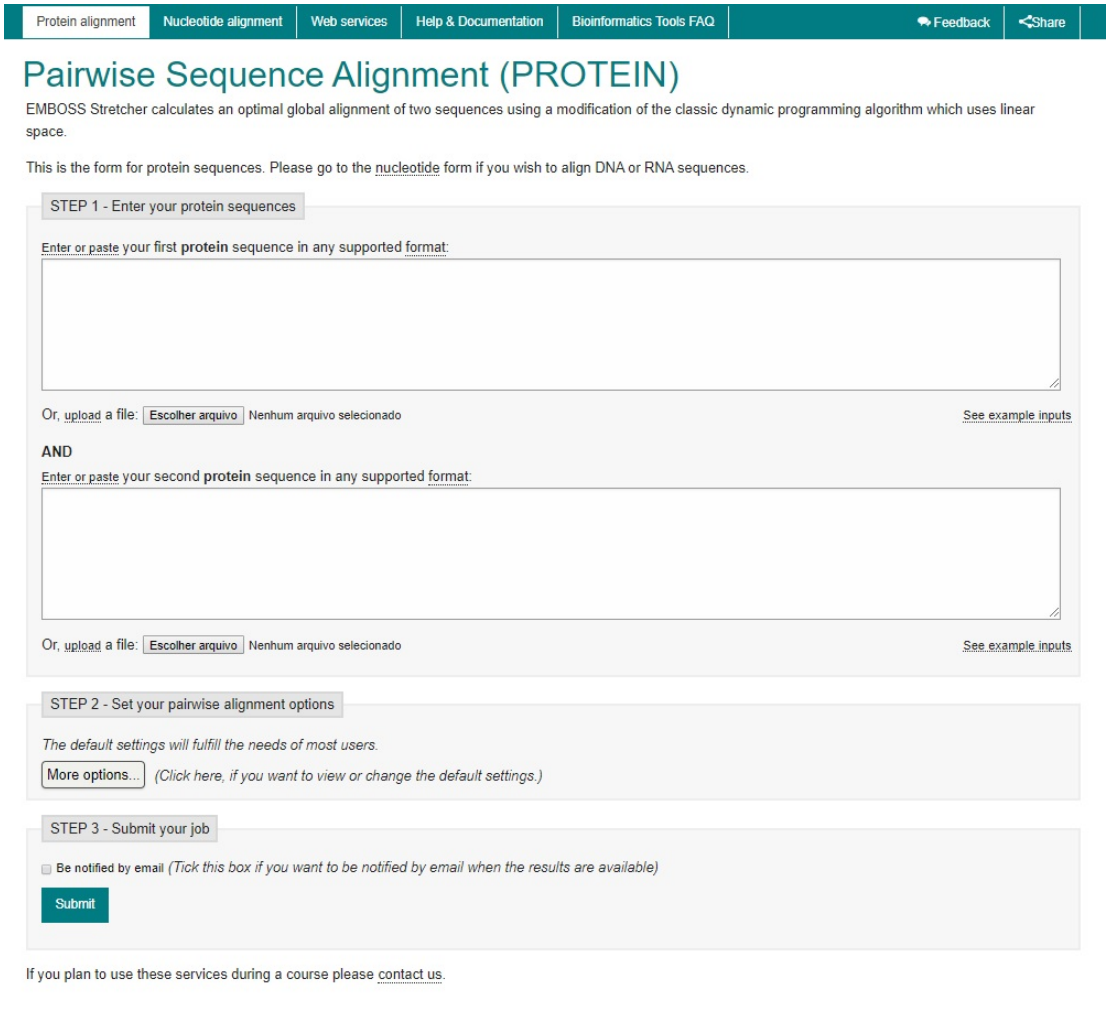

Figura 15: Entrada da ferramenta Emboss Matcher/Stretcher
Fonte: Extraído de https://www.ebi.ac.uk/Tools/psa/emboss_matcher

O Emboss Stretcher/Matcher é uma ferramenta conhecida por possuir um novo método de alinhamento, rápido e preciso, baseado no algoritmo de Needleman-Wunsch (Alinhamento Global) e Waterman-Smith (Local), utilizando uma estratégia de alinhamento progressivo somado ao algoritmo de alinhamento utilizado, sendo possível obter o valor do alinhamento ótimo usando apenas uma linha da matriz de programação dinâmica. Esse procedimento é mais adequado para uso com sequências maiores (MULLAN, 2006).

A figura 16 mostra o resultado gerado pela ferramenta Emboss Stretcher/Matcher.

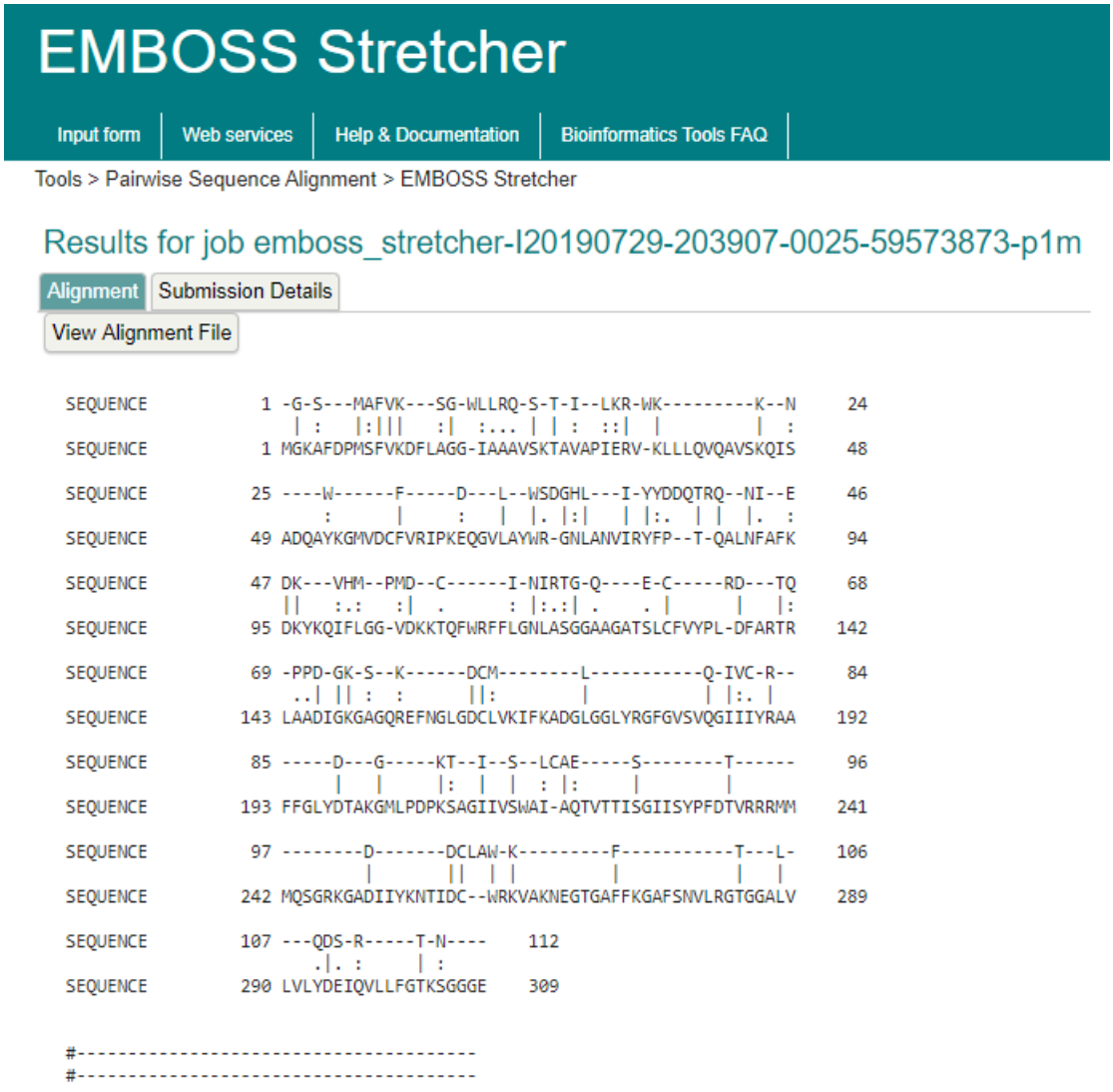

Figura 16: Saída da ferramenta Emboss Matcher/Stretcher

Fonte: Extraído de https://www.ebi.ac.uk/Tools/psa/emboss_matcher

3.3 ClustalW2

O ClustalW2 é uma ferramenta para a realização de alinhamento de sequências múltiplas de propósito geral para DNA ou proteínas. Assim como as outras ferramentas para alinhamento, o ClustalW2 calcula a melhor correspondência para as sequências selecionadas e as alinha para que as identidades, similaridades e diferenças possam ser observadas (LARKIN et al., 2007).

A figura 17 exibe a interface de entrada da ferramenta ClustalW2.

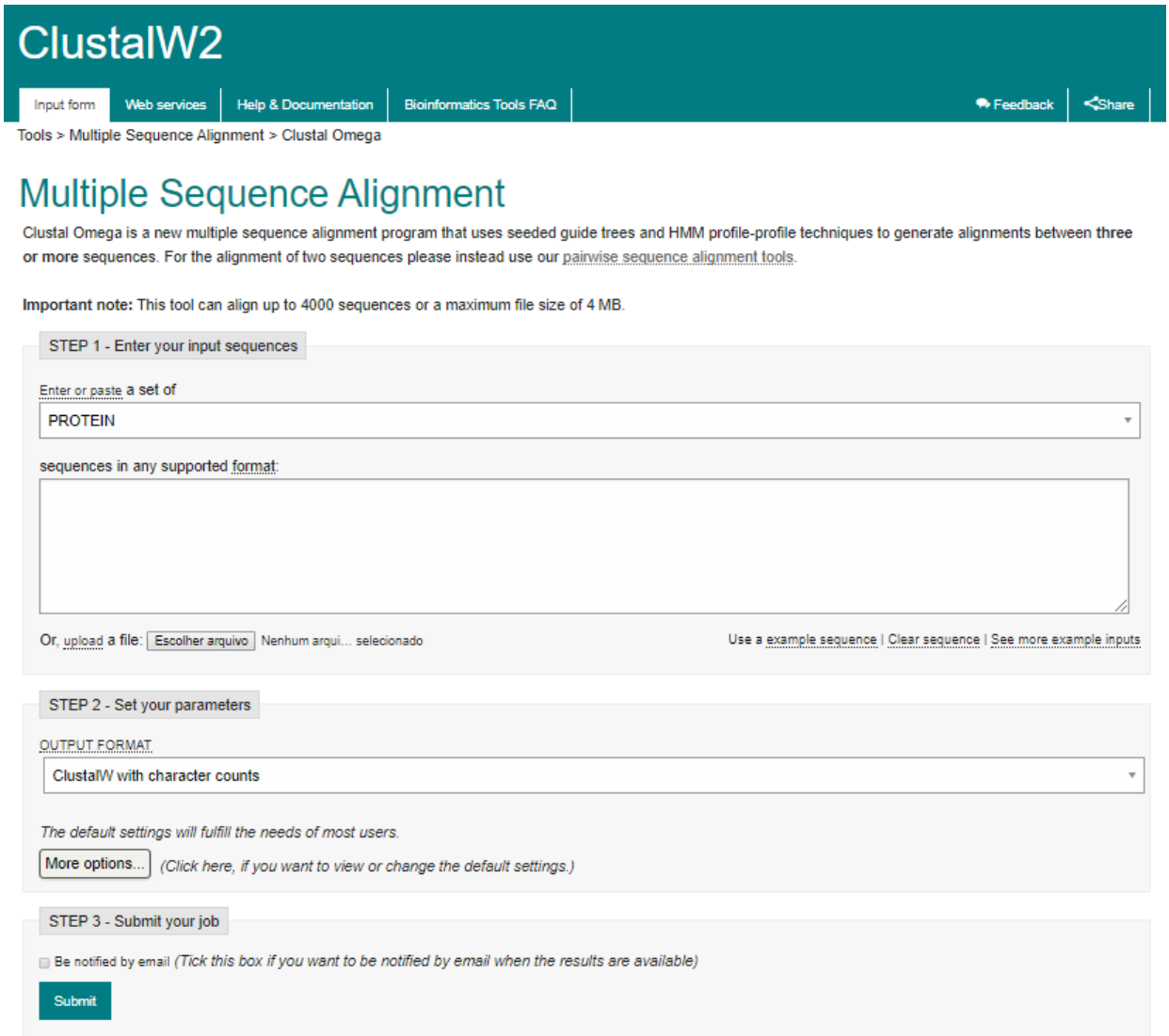

Figura 17: Entrada da ferramenta ClustalW2

Fonte: Extraído de http://www.clustal.org/clustal2

A execução é bem simples, primeiramente o usuário define as sequências que irão

servir de entrada na ferramenta, há a possibilidade do usuário alterar os parâmetros que já vem pré-estabelecido pela ferramenta. E, finalmente, o usuário define o título do trabalho e um endereço de e-mail para que o ClustalW2 se encarregue de exibir o resultado na tela e posteriormente enviar esse trabalho para o e-mail indicado (MCWILLIAM et al., 2013). Na figura 18 exibimos a saída gerada através da ferramenta ClustalW2.

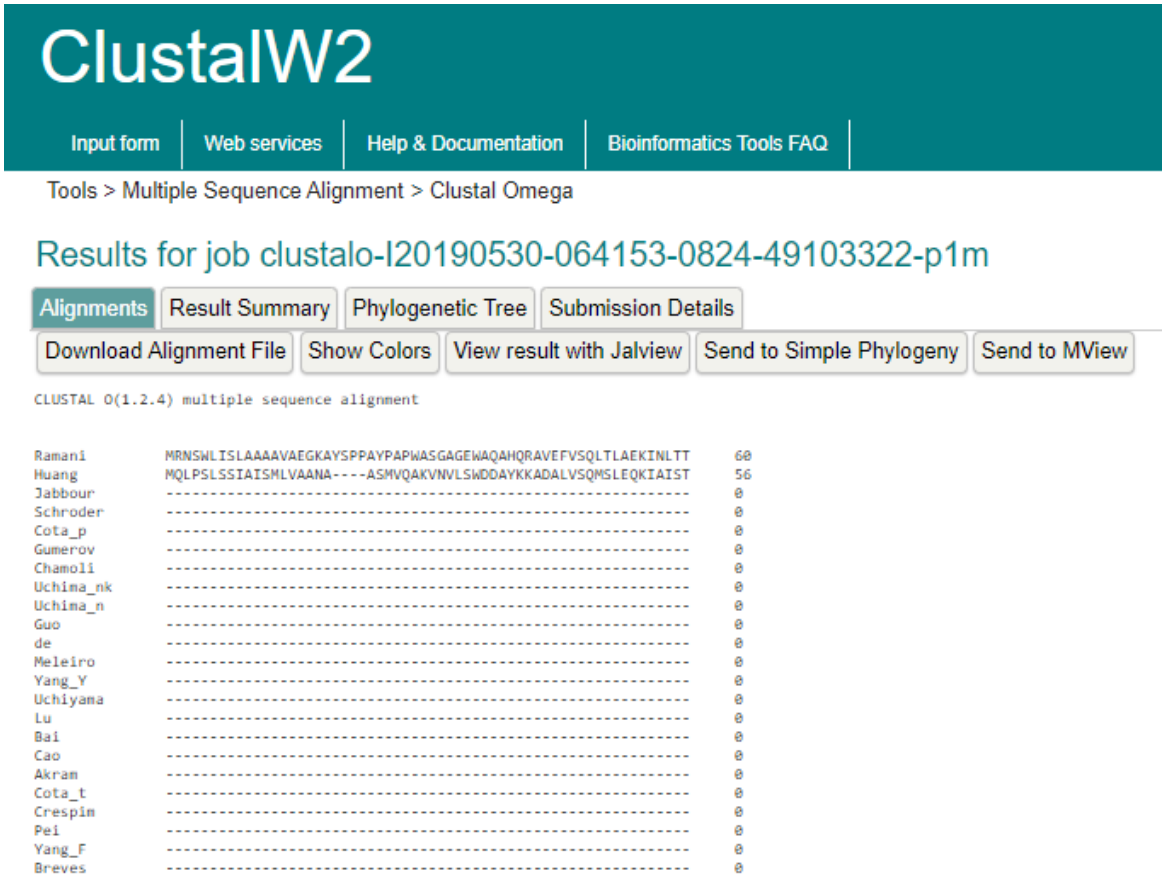

Figura 18: Saída da ferramenta ClustalW2

Fonte: Extraído de http://www.clustal.org/clustal2

3.4 Promals3D

O servidor web PROMALS3D executa alinhamentos múltiplos para sequências ou estruturas de proteínas utilizando um método progressivo que agrupa sequências similares (PEI; GRISHIN, 2014).

Além disso, o PROMALS3D melhora a qualidade de alinhamento de sequências distantemente relacionadas, combinando várias técnicas avançadas, como pesquisa de banco

de dados, predição de estrutura secundária e consistência probabilísticas de comparação. Proporciona aos pesquisadores uma ferramenta para produzir alinhamentos de alta qualidade e fidelidade (PEI; GRISHIN, 2007). Seu funcionamento consiste em 2 partes:

No primeiro estágio o Promals3D utiliza a função de pontuação da soma de pares (método BLOSUM62) - obtém como resultado uma série de grupos pré-alinhados que estão relativamente distantes um do outro.

No segundo estágio de alinhamento, uma sequência representativa é selecionada e são feitos buscas para recuperar homólogos do banco de dados UNIREF90[2] e PSIPRED[3]. Em seguida, um modelo de Markov oculto de alinhamentos de perfil-perfil com estruturas secundárias previstas é aplicado a pares de representantes para obter probabilidades posteriores de correspondências de resíduo. Abaixo é listado uma série de parâmetros no qual é utilizado na ferramenta Promals3D.

- PSIPRED pode prever a estrutura secundária de uma proteína (folhas-beta, alfa-hélices) da sequência primária.
- O banco de dados UNIREF90 combina sequências idênticas em uma única entrada
- Modelo de Markov, é um modelo estatístico que presta apoio aos problemas de decisão envolvendo incertezas em um período contínuo de tempo.
- BLOSUM62, é uma matriz de substituição utilizada no processo de alinhamento de sequências. Essa matriz é usada para pontuar o alinhamento entre as sequências diferentes, buscando regiões muito conservadas de famílias de proteínas e depois contam a frequência relativa de aminoácidos e as suas probabilidades de substituição.

O seu processo de alinhamento pode levar muito tempo para finalizar e depende diretamente do número de sequências divergentes utilizados. Uma forma de reduzir o tempo de execução é alterar o parâmetro “limite de identidade” (identify threshold), esse parâmetro é responsável por equilibrar a qualidade e a velocidade do alinhamento, porém alterando para executar de maneira mais rápida, poderá resultar um alinhamento menos preciso (PEI; KIM; GRISHIN, 2008). A figura 19 mostra a interface de entrada da ferramenta Promals3D.

[2]https://www.uniprot.org/help/uniref
[3]http://bioinf.cs.ucl.ac.uk/psipred/

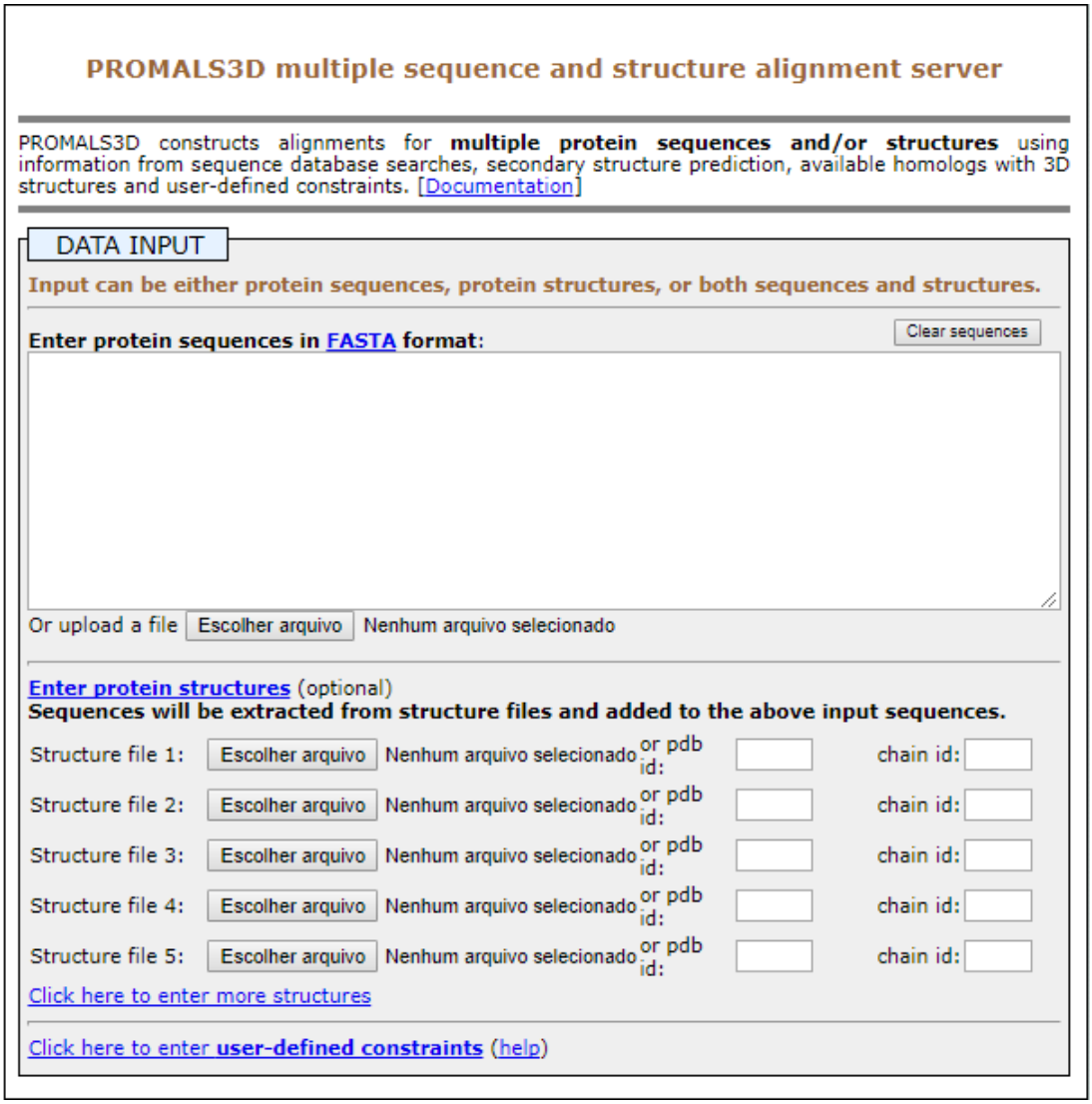

Figura 19: Entrada do Servidor PROMALS3D

Fonte: Extraído de http://prodata.swmed.edu/promals3d/promals3d.php

O Promals3D fornece links para mostrar o resultado dos alinhamentos realizados e possuem três formatos:

- COLORED: O agrupamento de sequências é refletido pela cor dos nomes de sequências. As estruturas secundárias previstas são mostradas para sequências representativas (os resíduos com fontes vermelhas e azuis são preditos para serem hélices α e feixes β, respectivamente);

- CLUSTAL: São relatadas informações úteis sobre o agrupamento das sequências, previsões de estrutura secundária, conservação de posição e sequência de consenso.

- FASTA: Mostra as sequências que foram inseridas no input e as sequências extraídas dos PDB's no formato FASTA para realizar o processo de alinhamento, em ordem de entrada.

Na figura 20 mostramos a o resultado da ferramenta Promals3D.

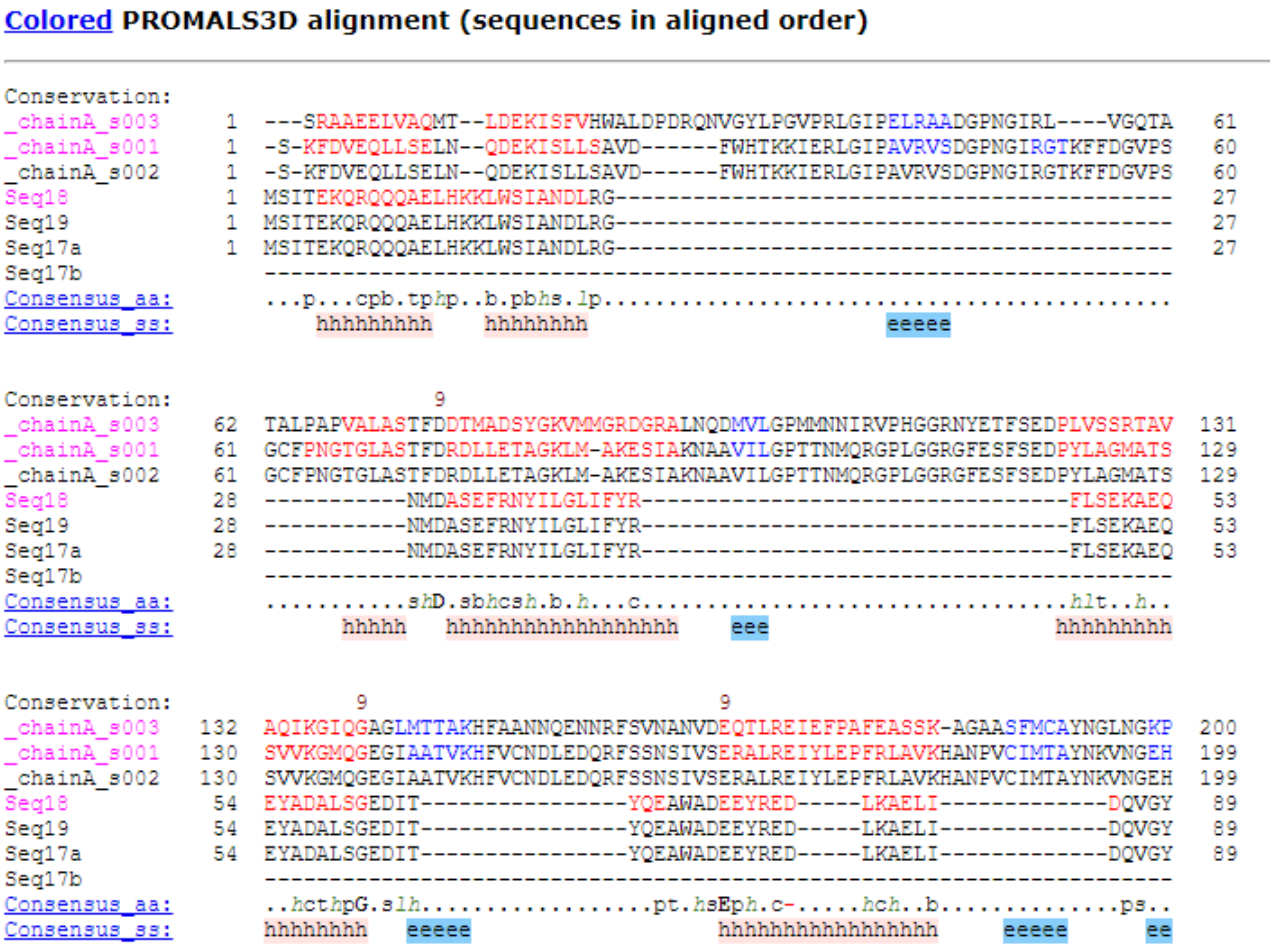

Figura 20: Resultado do Servidor PROMALS3D

Fonte: Extraído de http://prodata.swmed.edu/promals3d/promals3d.php

3.5 Clustal Omega

A ferramenta Clustal Omega é uma versão completamente reescrita e revisada da série de programas já existentes do Clustal para alinhamento de múltiplas sequências. Ele permite que centenas de milhares de sequências sejam alinhadas em apenas algumas horas, devido ao uso do algoritmo mBED para calcular a *guide trees*. Esse algoritmo permite

que problemas de alinhamento muito grandes sejam resolvidos rapidamente, mesmo em computadores pessoais. Além disso, a qualidade dos alinhamentos é superior às versões anteriores, conforme medido por uma série de referências populares, através do uso do método HHalign para alinhar os modelos ocultos de Markov no perfil. O programa atualmente é usado a partir da linha de comando ou pode ser executado on-line (SIEVERS; HIGGINS, 2014).

A figura 21 apresenta a interface de entrada da ferramenta Clustal Omega.

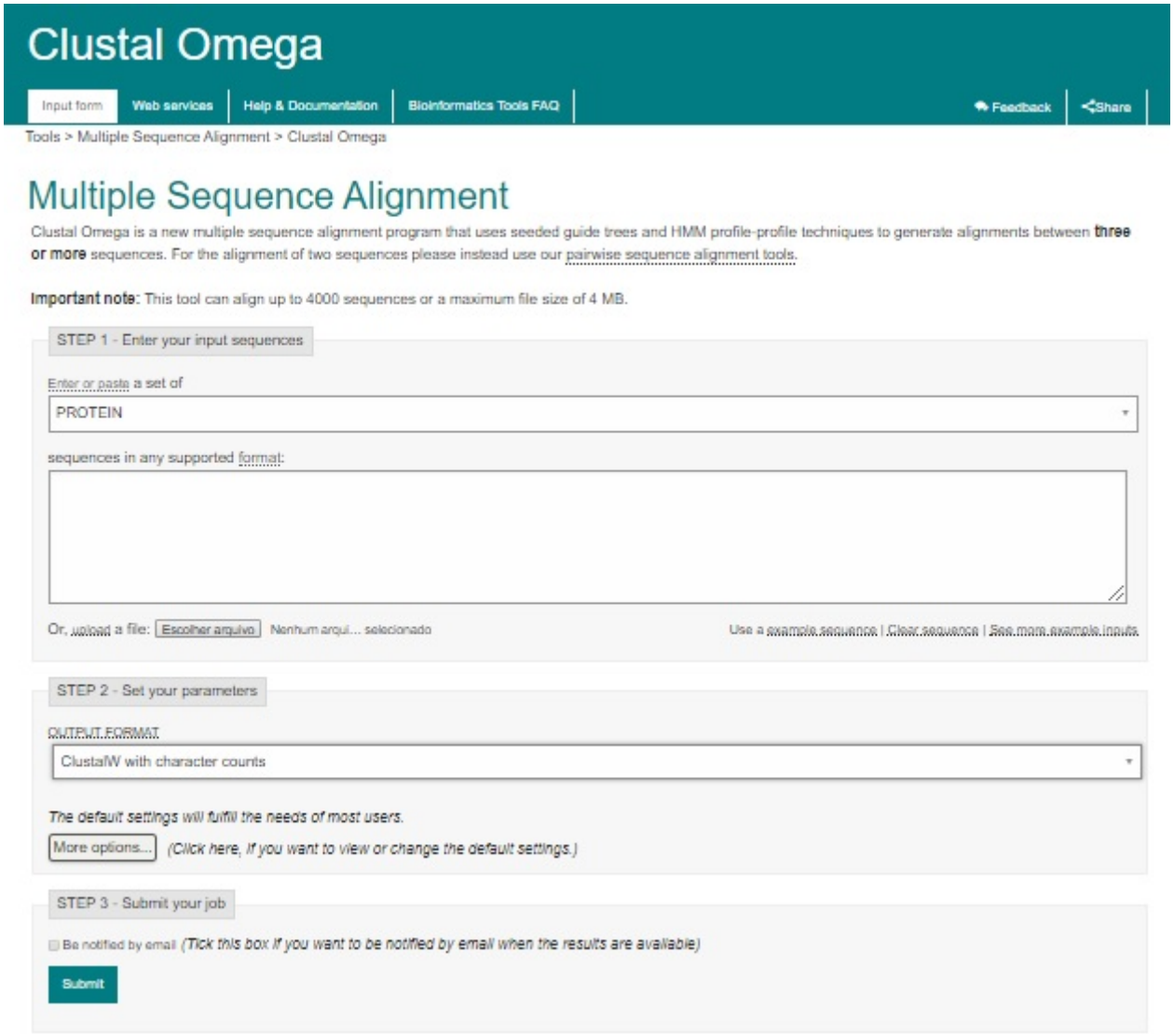

Figura 21: Clustal Omega

Fonte: Extraído de https://www.ebi.ac.uk/Tools/msa/clustalo

A figura 22 mostra o resultado gerado pela ferramenta Clustal Omega.

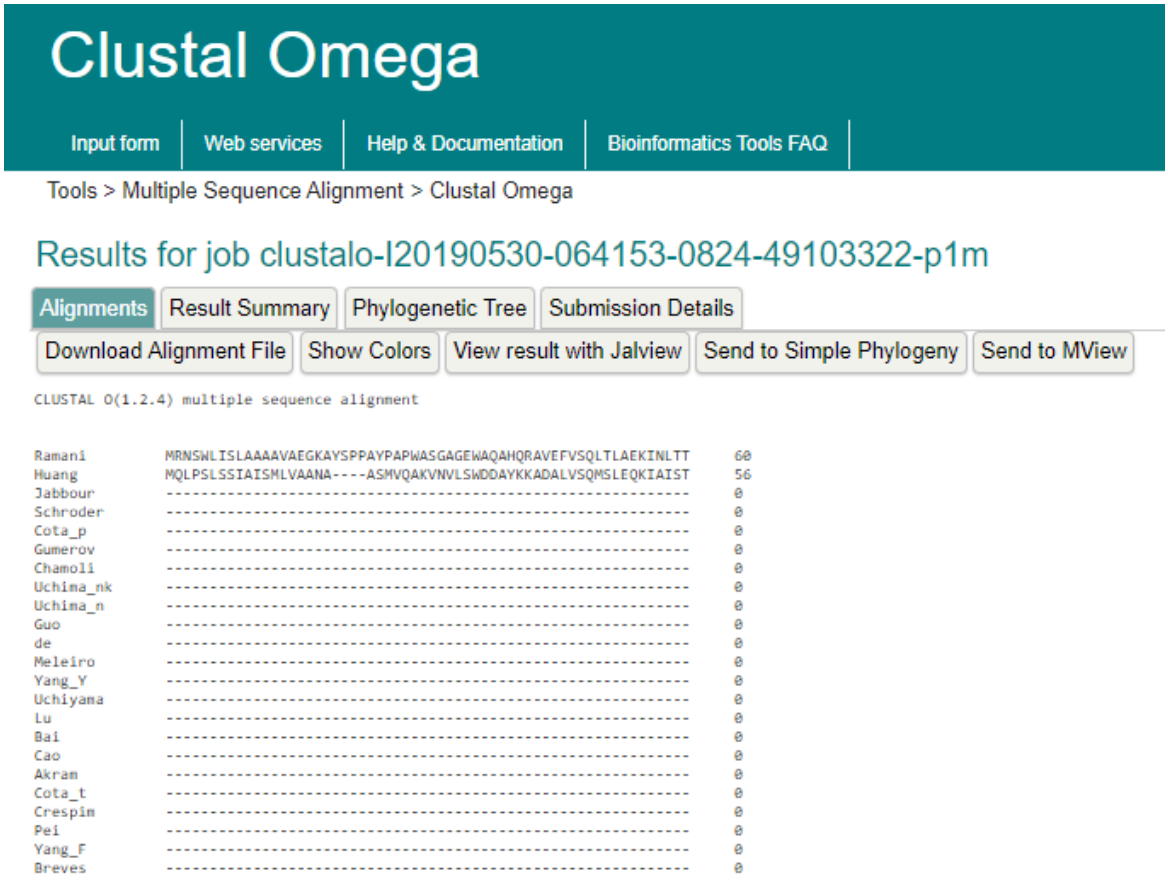

Figura 22: Resultado gerado pelo Clustal Omega

Fonte: Extraído de https://www.ebi.ac.uk/Tools/msa/clustalo

3.6 ICM Browser Pro

O ICM Browser é um produto gratuito da Molsoft e fornece a qualquer pesquisador o acesso direto informações relevantes da biologia estrutural e das famílias de proteínas, além de poder trabalhar com modelagem molecular e visualização e animação molecular de alto desempenho. Uma versão comercial, o ICM Browser Pro, possui recursos adicionais relacionados à criação, armazenamento e compartilhamento de informações estruturais, biológicas e químicas.

A firgura 23, mostra a interface do software ICM Browser Pro.

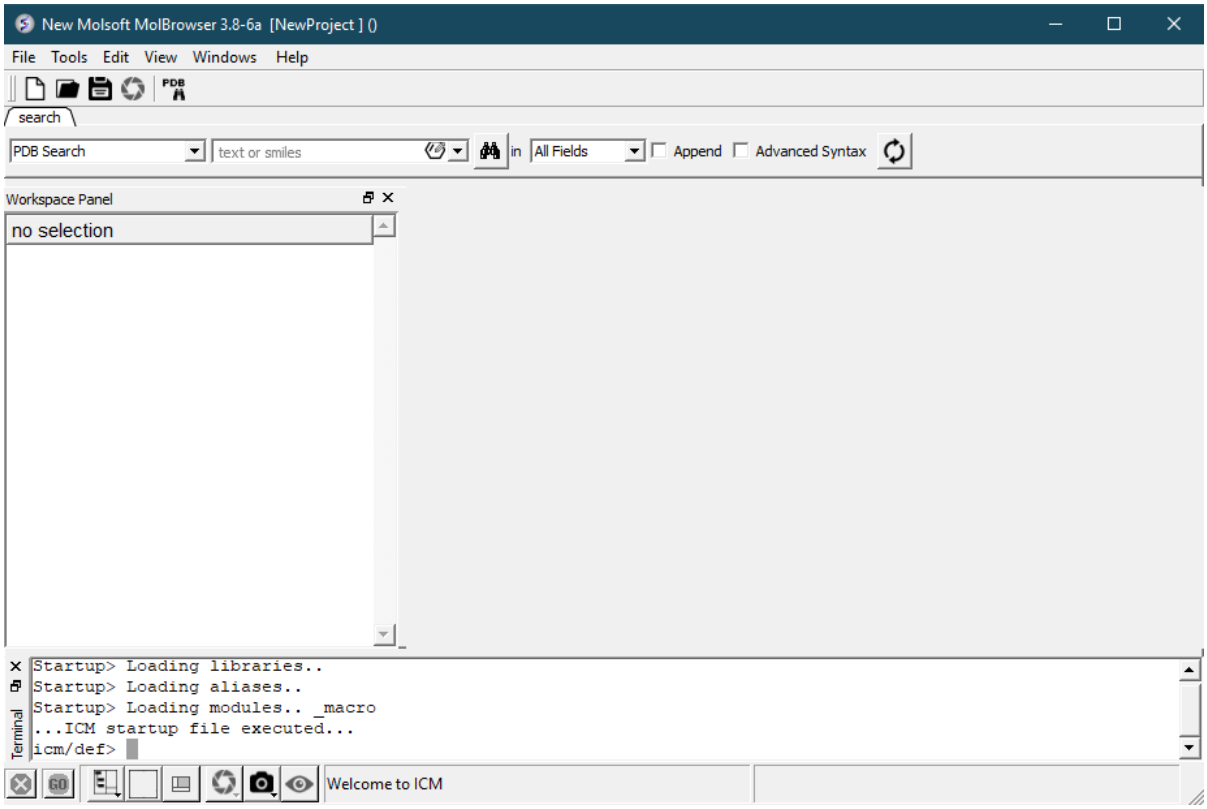

Figura 23: Interface do software ICM Browser Pro

Fonte: Extraído de http://www.molsoft.com

Esse software lê vários formatos de arquivo diretamente dos sites de banco de dados, incluindo o Protein Data Bank, propriedades físico-químicas, mapas de densidade eletrônica e arquivos de alinhamento e sequência. Além disso, fornece um rico ambiente gráfico molecular profissional com diferentes representações de proteínas, DNA e RNA, e múltiplos alinhamentos de sequências (ABAGYAN et al., 2004).

Na firgura 24, apresentamos o resultado gerado pela interface do software ICM Browser Pro.

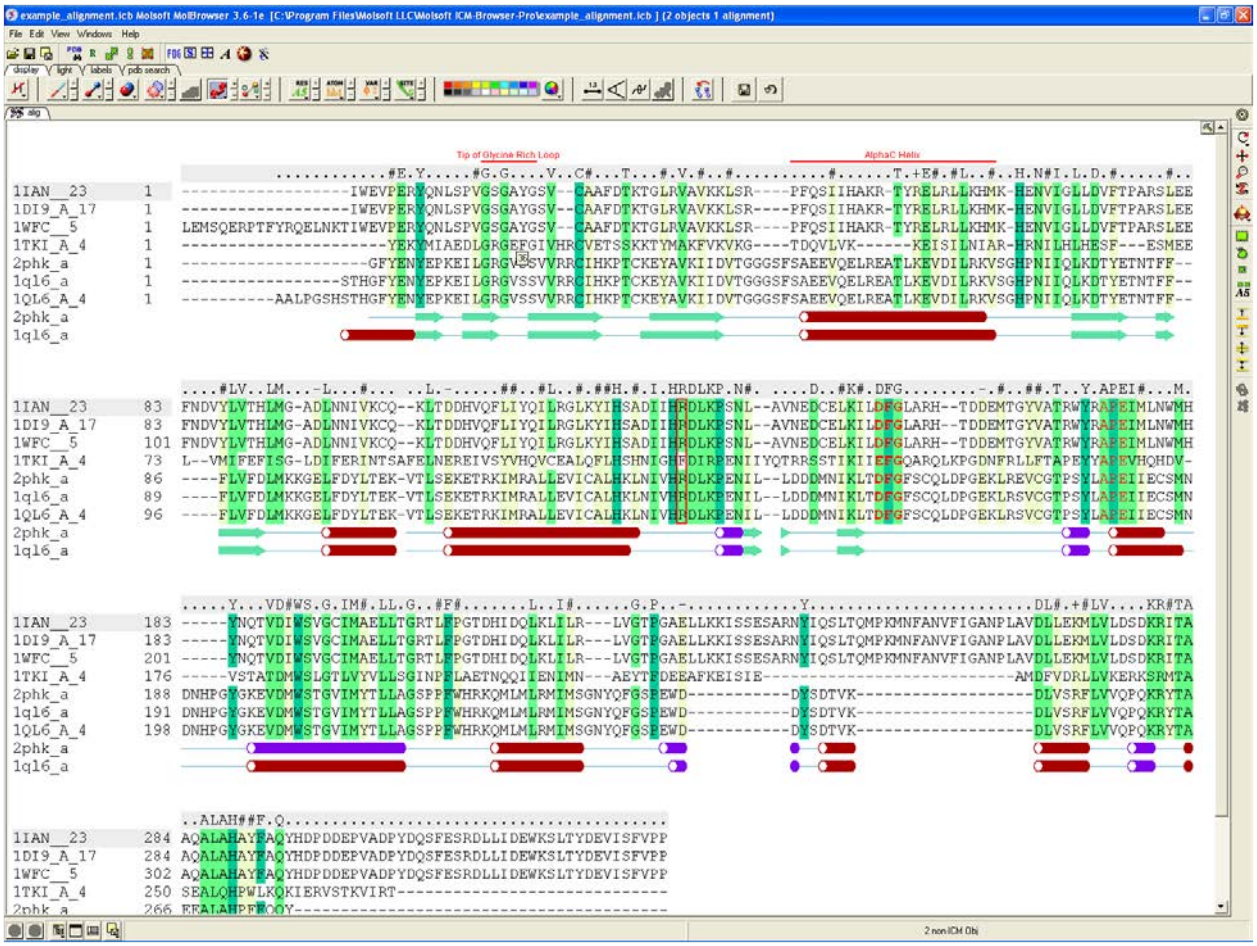

Figura 24: Resultado gerado pelo software ICM Browser Pro

Fonte: Extraído de http://www.molsoft.com

4 METODOLOGIA DE USO (ESSEX)

A metodologia de uso, está descrita na figura 25. O Essex permite ao usuário a liberdade para que ele possa inserir suas sequências de diversas origens e experimentos diferentes.

As sequências fasta podem ser obtidas em diversos bancos de dados como NCBI Protein, Uniprot ou RSCB Protein Data Bank, onde deve ser realizado o *download* da sequência no computador e posteriormente inserir a sequência no Essex através do botão *Upload*, também é possível digitar ou copiar a sequência por meio de uma caixa de texto que a ferramenta disponibiliza.

Além disso, é possível inserir a estrutura da proteína por intermédio do botão de *Upload* ou ser obtida no banco de dado RSCB Protein Data Bank onde a ferramenta se responsabiliza de buscar e extrair a sequência de aminoácidos referente a estrutura tridimensional da proteína através de um *script* de Biojava.

Escolhido os dados de entrada da ferramenta, o usuário deverá escolher nessa interface qual é a sequência de referência e posteriormente qual tipo alinhamento a ser realizado, dentre elas é possível realizar o alinhamento local, global ou múltiplo.

O resultado após a realização do alinhamento é que a ferramenta tenha como saída sequências alinhadas corretamente e a exibição dessas informações de maneira clara e objetiva, através da representação visual somado com o esquema de cor e a numeração correta para cada aminoácido, fazendo com que o usuário localize com mais clareza o aminoácido de interesse nas sequências que foram alinhadas.

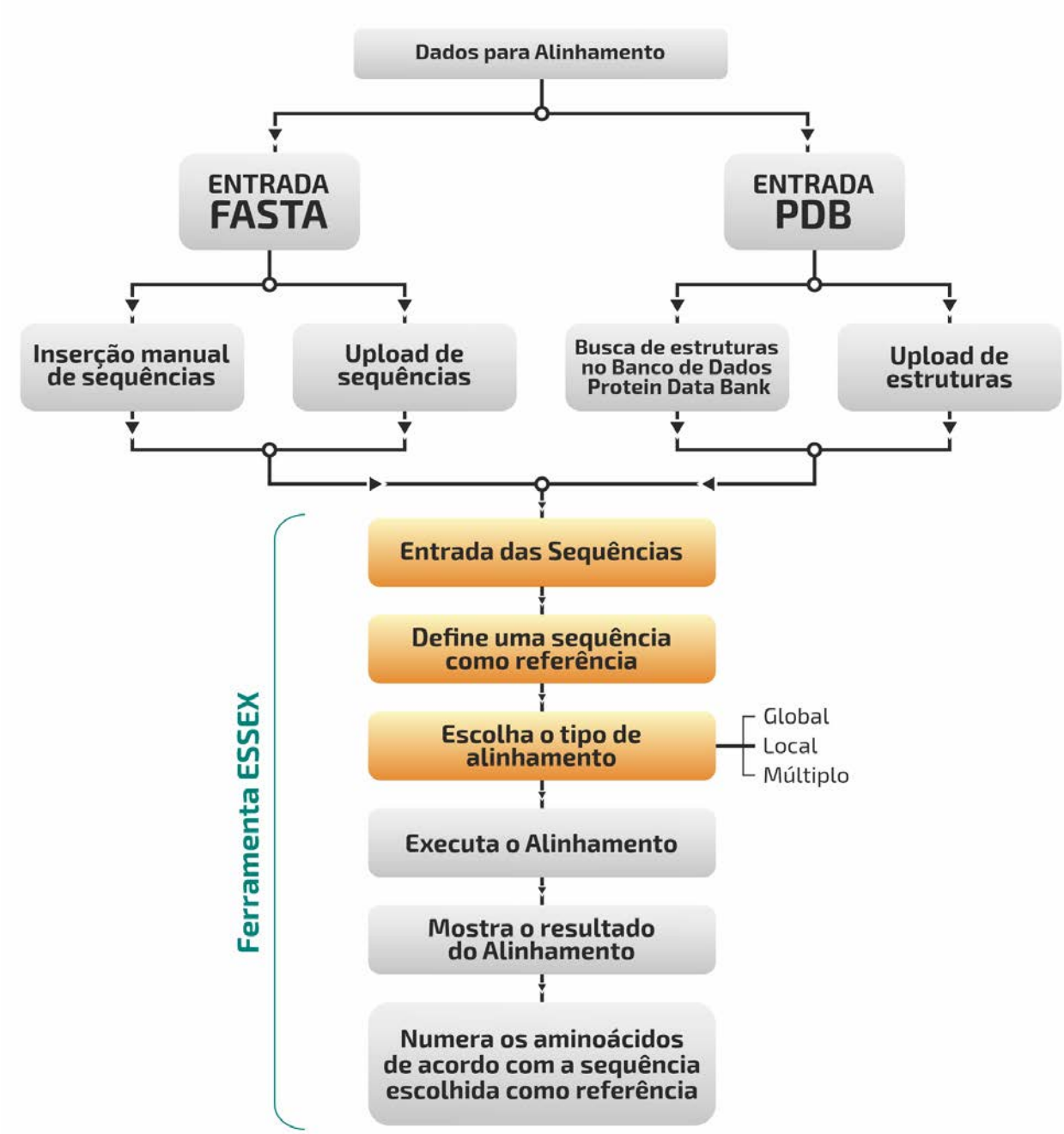

Figura 25: Representação da Metodologia Proposta.

Fonte: Dados do Autor.

4.1 Ferramentas Utilizadas

Para o desenvolvimento do Essex foram utilizadas as seguintes ferramentas:

- BIOJAVA: O BioJava foi criado em 1999 por Thomas Down e Matthew Pocock como uma interface de programação de aplicativo (API) para simplificar o desen-

volvimento de software de bioinformática usando Java (POCOCK; DOWN; HUBBARD, 2000). Consiste em um projeto de código aberto que fornece ferramentas Java para execução de processos biológicos, tendo como objetivo, automatizar e otimizar tarefas rotineiras na área de bioinformática. É um projeto aberto com muitos usuários e suporte. Possui uma biblioteca de ferramentas para tarefas comuns de bioinformática. A página do BioJava[1] fornece acesso ao código-fonte e à documentação detalhada para que bioinformatas possam utilizar (HOLLAND et al., 2008).

- BOOTSTRAP: È uma ferramenta gratuita que serve para desenvolver componentes de interface para aplicações web e sites, usando HTML, CSS e JavaScript. Através de sua biblioteca, também é possível criar projetos responsivos para dispositivos móveis. O Bootstrap, é uma das ferramentas mais importantes para o desenvolvimento de sites. Isso é dado pela sua simplicidade e também porque seus padrões seguem os princípios de usabilidade e as tendências de design para interfaces (OTTO; THORNTON et al., 2015).

- JAVASCRIPT: È uma linguagem de programação utilizado no front-end, utilizado para controlar o HTML e o CSS para manipular o comportamento na página. Trabalhando paralelamente com o bootstrap, o Javascript vem com vários componentes no formulário de plugins jQuery. Eles fornecem mais elementos de interface do usuário, tais como caixas de diálogo, dicas, carrosséis, botões e tooltip (FLANAGAN, 2006).

- MHTML: É uma extensão para um formato de arquivos de Página Web salvo pelo navegador. O MHTML salva o conteúdo da página da web e incorpora recursos externos, como imagens, animações e assim por diante, em documentos HTML. Logo, todos os links relativos do documentos HTML serão remapeados para que o conteúdo possa ser localizado, de maneira que o usuário possa abrir uma página web sem a conexão de internet e não perder a dinâmica do conteúdo da página (PALME;

[1] https://biojava.org/

HOPMANN; SHELNESS, 1999).

- APACHE PDFBOX: O Apache PDFBox é uma biblioteca Java de Software Livre, por isso é fácil de usar com uma ampla variedade de linguagens de programação, incluindo Java, Groovy, Scala, Clojure, Kotlin e Ceylon. Ele serve para trabalhar com documentos PDF que permite a criação de documentos PDF, a manipulação de documentos existentes e a capacidade de extrair conteúdo de documentos (PDFBOX, 2014).

- APACHE TOMCAT: O software Tomcat, desenvolvido pela Fundação Apache, permite a execução de aplicações para web de código aberto. Sua principal característica técnica é estar centrada na linguagem de programação Java, mais especificamente nas tecnologias de Servlets e de Java Server Pages (JSP). Essa é um dos muitos produtos de código aberto relacionados ao Apache Software Foundation usados por profissionais de TI para várias tarefas e objetivos (VUKOTIC; GOODWILL, 2011).

- JRE (Java Runtime Environment): Consiste no Java Virtual Machine (JVM), nas classes centrais e bibliotecas de suporte da plataforma Java. Ele representa a parte responsável pelo tempo de execução do software Java (VENNERS, 1998). Com isso e o APACHE TOMCAT seria praticamente tudo de que você precisa para executar a ferramenta Essex em um navegador web do seu computador.

4.2 Ferramenta Proposta

O Essex é executado em um ambiente web e foi desenvolvido para realizar alinhamentos de sequências de proteínas e/ou estruturas, no intuito de renumerar cada aminoácido da sequência no processo de alinhamento e consequentemente para que o biólogo ou qualquer usuário da ferramenta localize o aminoácido pretendido de maneira fácil e rápida. A partir da sua interface o usuário define o seu conjunto específico de sequências de proteínas, essas sequências por sua vez, podem estar armazenados no seu computador ou através de palavras chaves no campo de busca da própria ferramenta. Esse processo de

busca de sequências por meios de palavras chaves é dado através do uso da internet em conjunto com o *RCSB PDB RESTful Web Service interface* que utiliza o banco de dados do *Protein Data Bank*. Para compreender esse tipo de serviço, ele tem a finalidade de trocar informações entre duas entidades de software através da internet, utilizando os protocolos de comunicação disponíveis, fazendo com que seja realizado a aquisição de informações do PDB no Essex.

A partir desse momento, sem mais a utilização da conexão com a internet, o Essex realiza alinhamento global, local ou múltiplo de maneira que numere automaticamente cada aminoácido correspondente na sequência, mostrando sua real posição e sua posição após o processo de alinhamento.

Múltiplos uploads de sequências de proteínas e/ou estruturas podem ser inseridas no Essex, com base da escolha do usuário o programa gera um alinhamento global, local ou múltiplo a partir das sequências que foram inseridas. Fazendo que a ferramenta tenha esse fator positivo, permitindo que o usuário desta ferramenta tenha o livre-arbítrio de inserir sequências de origens e locais distintos sem nenhuma restrição e limitação. 26 exibe a interface da ferramenta.

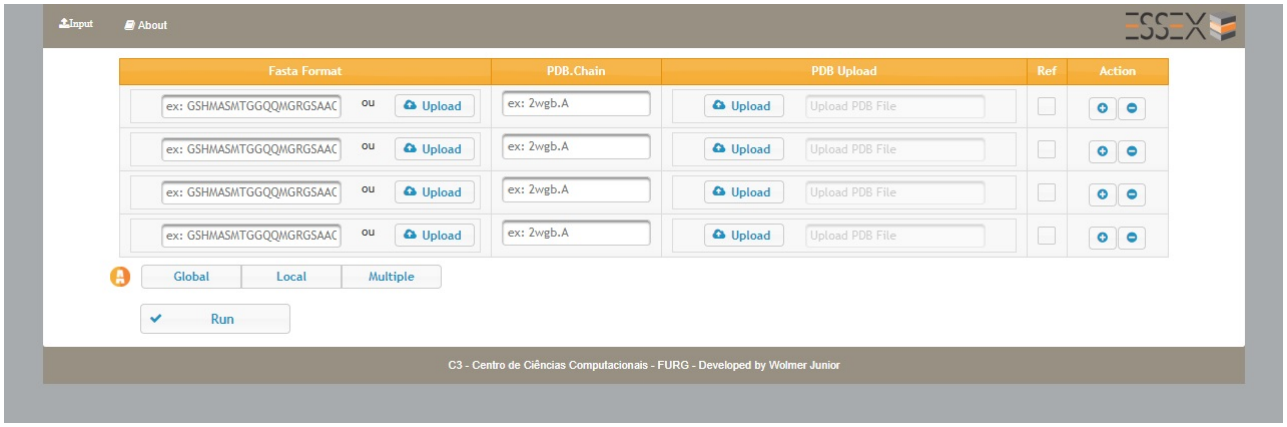

Figura 26: Interface da ferramenta Essex

Fonte: Dados do Autor.

4.2.1 Entrada

Os usuários da ferramenta podem inserir ou fazer *upload* de sequências de proteínas no formato fasta, ou realizar estruturas de sequências no banco de dados (RCSB Protein

Data Bank) que contém uma grande quantidade de informação simplesmente introduzindo o ID da estrutura.

Também é possível que o usuário carregue arquivos estruturais, essas estruturas devem estar no formato PDB e o Essex extrai a sequência da estrutura. A saída fica por conta da escolha do usuário, podendo optar pela realização do alinhamento global, local ou múltiplo.

A ferramenta pode levar um tempo considerável para concluir o processo de alinhamento devido ao grande número de entradas de sequências divergentes, o tempo é variável e é de acordo com as características e configuração de cada computador. Por fim, é importante mencionar que não existe uma limitação de sequências que possam ser inseridas e basta clicar em um botão para adicionar ou remover uma sequência.

4.2.2 Saída

A ferramenta Essex apresenta uma saída que é dada a partir do método de alinhamento que o usuário selecionou.

- **Alinhamento Global:** Utiliza o algoritmo de Needleman Wunsch para esse propósito e pode usar apenas duas sequências para o alinhamento, caso usuário tenha escolhido mais de duas sequência e determinando uma delas como referência, o Essex dividirá em abas os respectivos alinhamentos para melhor entendimento, sempre mantendo a sequência definida como referência em todas as abas e as demais sequências alinhadas com a referência.

- **Alinhamento Local:** O processo é bem parecido com o alinhamento global, o que diferencia é que o programa apresenta somente a região que obteve o maior score de alinhamento. Nesse procedimento é utilizado o algoritmo de Smith-Waterman

- **Alinhamento Múltiplo:** No método de alinhamento múltiplo é utilizado o algoritmo ClustalW que serve para o alinhamento de três ou mais sequências e é amplamente utilizado pois é computacionalmente eficiente e precisa, produzindo alinhamentos de sequências múltiplas biologicamente significativas de sequências divergentes (CHENNA et al., 2003).

Na figura 27 é exibido um exemplo de resultado de um processo de alinhamento múltiplo realizado pelo Essex, onde foram inseridas uma sequência fasta e quatro estruturas tridimensionais utilizando o botão de *Upload*.

Figura 27: Exemplo de resultado de alinhamento múltiplo realizado pelo Essex.

Fonte: Dados do Autor.

A ferramenta Essex possibilita que o usuário exporte seus resultados de alinhamentos nos seguintes formatos: TXT, HTML e PDF através dos respectivos botões *Donwload TXT*, *Download HTML* e *Download PDF*. A ferramenta oferece um botão *toggle* denominado como *Subtitles* que tem a função de mostrar e esconder a informação de legenda que informa a propriedade de cada aminoácido.

Um outro recurso que a ferramenta dispõe para auxiliar o usuário a localizar o aminoácido pretendido em um processo de alinhamento é a utilização de *Tooltips*, que baseia-se em um balão informativo que permite o usuário saber a posição real e a posição que ficou após o processo de alinhamento. Essa informação é exibida quando o usuário passar o mouse no aminoácido, como é exibido na figura 28

Figura 28: Balões informativos da ferramenta ESSEX

Fonte: Dados do Autor.

Além da funcionalidade de *Tooltip* o Essex disponibiliza da função *Popups*, que baseia em um resumo da posição real e posição final do aminoácido pretendido em todas sequências que foram inseridas na ferramenta Essex. Para obter essa informação o usuário deverá clicar no aminoácido, como mostra na figura 29.

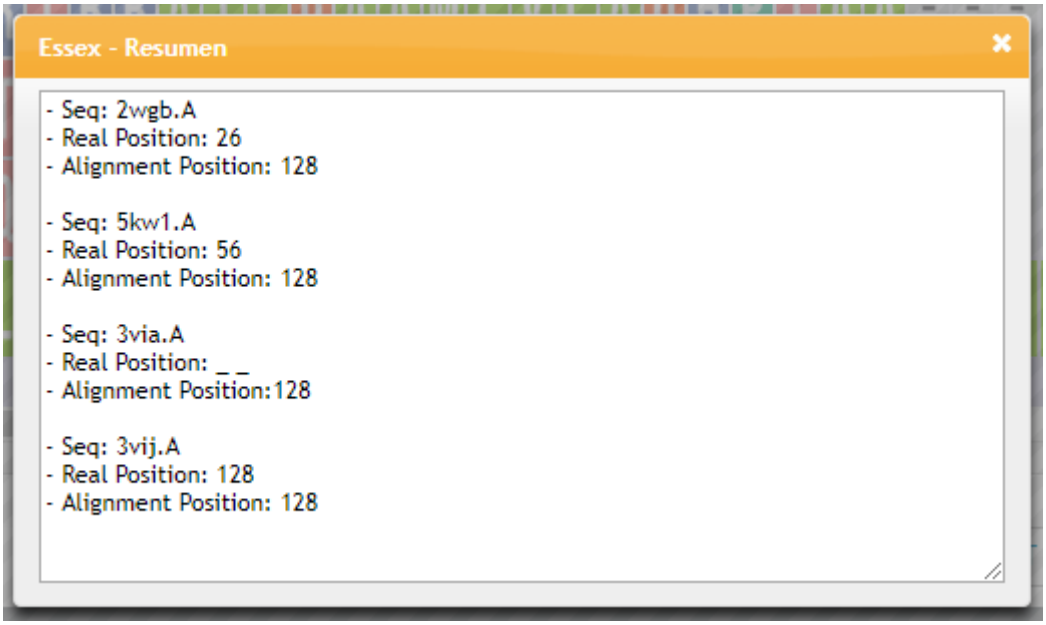

Figura 29: Popup's da ferramenta ESSEX

Fonte: Dados do Autor.

4.2.2.1 Esquema de Cores

Após o resultado de alinhamento e a numeração de cada aminoácido realizado com sucesso, a exibição dessas informações leva um esquema de cores baseado no CINEMA, utilizamos esse esquema de cores já existente por obter mais informações das proprieda-

des do aminoácido.

Além de informar a propriedade de cada aminoácido, auxilia o usuário desta ferramenta a localizar facilmente o aminoácido de interesse nas sequências que foram alinhadas (RASMOL, 2002). A figura 30 representa os esquemas de cores que são utilizados por ferramentas que realizam o processo de alinhamento.

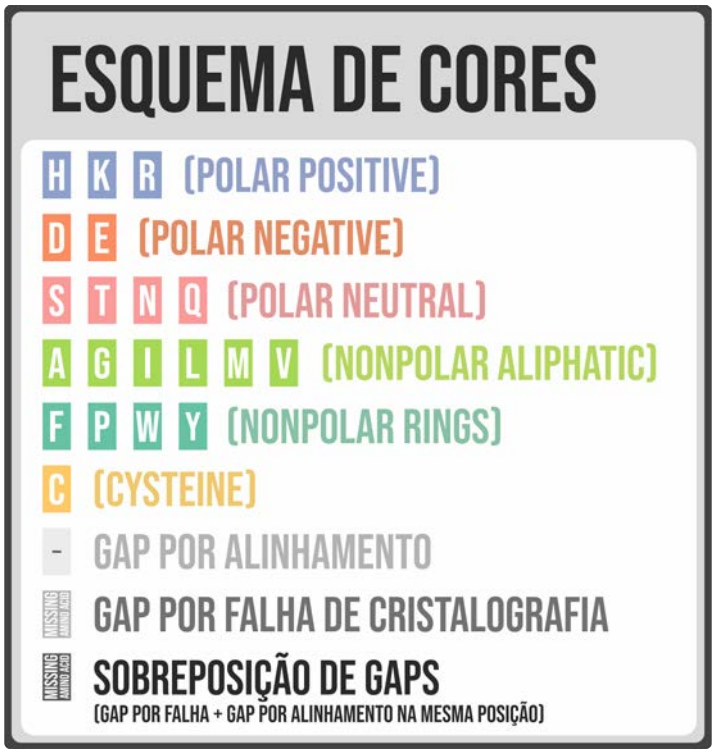

Figura 30: Representação do Esquema de Cores da Ferramenta Essex.

Fonte: Dados do Autor.

4.2.3 Funcionalidades

Essa seção apresenta as funcionalidades da ferramenta. Para esse fim, utilizaremos para a realização desse procedimento uma sequência teste do tipo fasta: 3VIJ

4.2.3.1 Funcionalidade 1: fasta x fasta

Nesse momento, são submetidos dois arquivos no formato fasta: 3VIJ e 3VIK para a realização do alinhamento. Conforme é mostrado na figura 31 é possível ver que tratam de sequências idênticas e completas e por isso são alinhadas de ponta a ponta sem a utilização de *GAP's*. É importante ser mencionado que em todos os processos de alinhamento,

possuem campos de dicas da posição real e da posição de alinhamento de cada aminoácido *(tooltips)*.

Suponhamos que há interesse no aminoácido Leucina (L) que fica situado na posição 18 da sequência de referência, por se tratarem de duas sequências idênticas é notório que a informação de *Real Position* e *Alignment Position* haverá o mesmo valor.

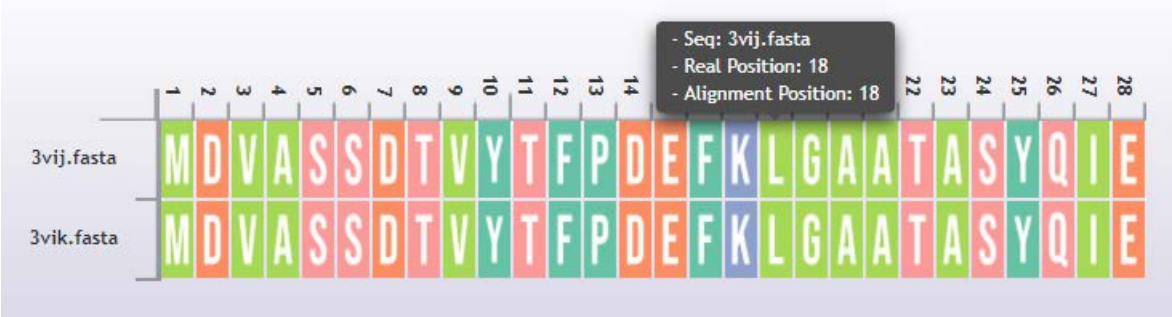

Figura 31: Teste realizado: fasta 3VIJ x fasta 3VIK

Fonte: Dados do Autor.

4.2.3.2 Funcionalidade 2: fasta 3VIJ x fasta Q8T0W7

Na figura 32 apresenta o caso onde é inserido um arquivo no formato fasta da proteína: 3VIJ e a outra sequência é a sequência Q8T0W7 do banco de dados Uniprot que tem como identidade 92,2% comparada ao 3VIJ, isso foi realizado para que possamos observar a inserção de gaps de alinhamento que o Essex realiza para obter o melhor alinhamento entre essas duas sequências.

No mesmo contexto da funcionalidade anterior, em que temos o interesse de localizar o aminoácido da posição 18 da sequência de referência em outra sequência. Neste caso, por tratar de sequências com identidades diferentes, podemos perceber que depois do processo de alinhamento a Leucina (L) permanece na posição 18 na sequência de referência, porém na segunda sequência é encontrada na posição *(Real Position)* número 37.

Figura 32: Teste realizado: fasta 3VIJx fasta Q8T0W7

Fonte: Dados do Autor.

4.2.3.3 Funcionalidade 3: fasta x PDB

É submetido o *upload* da proteína e extraído a coluna *ATOM* como primeira sequência e o formato fasta da mesma proteína (3VIJ) como a segunda sequência realizado no teste. Esse exemplo mostra que existe a inserções dos *missings amino acids* na cor cinza e na cor preta.

Para casos de sequências que apresentam problemas relacionados aos experimentos utilizados para gerar as estruturas tridimensionais eles possuem a cor cinza e as inserções dos *missings amino acids* na cor preta é dada quando há sobreposição de *gaps* de alinhamento somado ao *gap* por falha de cristalografia, como é mostrada na figura 33. Tornando uma das principais vantagens que o Essex disponibiliza comparado aos outros softwares relacionados.

Ao localizar a Leucina (L) que ocupa a posição 18 da sequência de referência, na outra sequência que está no formato PDB da proteína 3VIJ, é perceptível a presença de *Missings Amino Acids* uma vez que é comum conter aminoácidos faltantes na coluna ATOM no arquivo do formato PDB da proteína. Nessa funcionalidade, a Leucina (L) é localizada na posição 37.

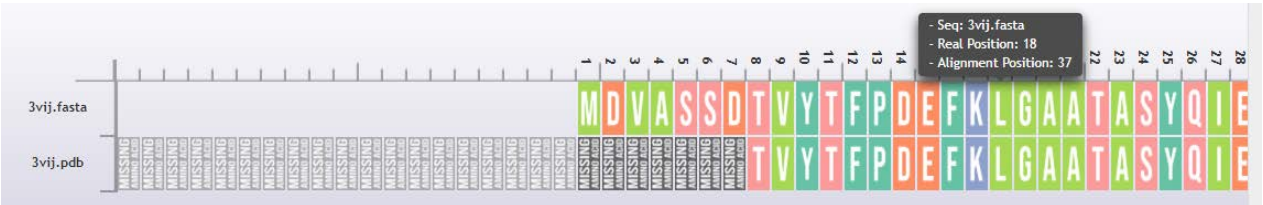

Figura 33: Teste realizado: PDB x fasta da proteína de código PDB: 3VIJ

Fonte: Dados do Autor.

4.2.3.4 Funcionalidade 4: PDB x PDB (diferentes)

Na figura 34, é notável considerar que a inserções dos *missings amino acids* não influenciam no resultado do alinhamento, eles apenas situam onde aparecem as falhas.

Figura 34: Teste realizado: PDB das proteínas de código PDB: 2WGB x 3VIJ

Fonte: Dados do Autor.

4.3 Comparação entre os Trabalhos Relacionados

Na figura 35, mostramos um comparativos sobre os trabalhos relacionados: Promals3D, Vermont, Emboss Stretcher, Emboss Matcher, Clustal W2, Clustal Omega e ICM Browser Pro com a ferramenta Proposta: Essex. Apresentamos na tabela comparativa algumas funcionalidades de cada ferramenta e destacamos que existem funcionalidades que somente o Essex realiza, como exemplo, exibir um resumo referente as posições de cada aminoácido ou a visualização dos aminoácidos faltantes no processo de cristalografia.

	PROMALS3D	VERMONT	EMBOSS STRETCHER	EMBOSS MATCHER	CLUSTAL W2	CLUSTAL OMEGA	ICM BROWSER PRO	ESSEX
PERMITE UPLOAD DE ARQUIVOS DE EXTENSÃO .PDB DO COMPUTADOR DO USUÁRIO	✓	✗	✓	✓	✓	✓	✓	✓
INSERÇÃO DE FASTA FILE PARA O ALINHAMENTO	✓	✓	✓	✓	✓	✓	✓	✓
REALIZA O ALINHAMENTO MÚLTIPLO	✓	✓	✗	✗	✓	✓	✓	✓
REALIZA O ALINHAMENTO LOCAL	✗	✗	✗	✓	✗	✗	✗	✓
REALIZA O ALINHAMENTO GLOBAL	✗	✗	✓	✗	✗	✗	✓	✓
INSERÇÃO PDB E CADEIA PARA BUSCA NO PROTEIN DATA BANK	✓	✓	✗	✗	✗	✗	✓	✓
MÚLTIPLOS UPLOADS DE ARQUIVOS PDB	✓	✗	✗	✗	✗	✗	✓	✓
MÚLTIPLAS INSERÇÕES DE FASTA FILE	✗	✗	✗	✗	✓	✓	✓	✓
EXIBE UM RESUMO, REFERENTE AS POSIÇÕES CLICANDO NO AA PRETENDIDO	✗	✗	✗	✗	✗	✗	✗	✓
VISUALIZAÇÃO DOS AMINOÁCIDOS FALTANTES NO PROCESSO DE CRISTALOGRAFIA NO ARQUIVO PDB	✗	✗	✗	✗	✗	✗	✗	✓
BALÕES INFORMATIVOS MOSTRANDO A POSIÇÃO REAL E A POSIÇÃO DO ALINHAMENTO DO AA	✗	✓	✗	✗	✗	✗	✗	✓
PERMITE O DOWNLOAD EM HTML	✗	✗	✗	✗	✗	✗	✗	✓
PERMITE O DOWNLOAD EM PDF	✓	✗	✗	✗	✗	✗	✗	✓
PERMITE O DOWNLOAD EM TXT	✓	✗	✗	✗	✗	✗	✓	✓

Figura 35: Comparação Entre as Ferramentas
Fonte: Dados do Autor.

5 ESTUDO DE CASO

Essa seção tem como objetivo de comparar o uso de ferramentas distintas que realizam alinhamentos e validar o alinhamento realizado pelo Essex. As três ferramentas utilizadas: Essex, Clustal Omega e Promals3D foram escolhidas por serem utilizadas nos artigos citados.

5.1 Estudo de Caso 1

Esse primeiro caso, é baseado no alinhamento que foi realizado no estudo sobre os efeitos de nanotubos de carbonos em mecanismos mitocondriais e correlações *in silico* estrutura-atividade.

Em (GONZÁLEZ-DURRUTHY et al., 2017) são avaliados os efeitos induzidos por uma família de nanotubos de carbono sobre mecanismos mitocondriais chave, através da integração de metodologias *in vitro* e *in silico*, baseadas na predição de relações quantitativas estrutura-atividade. Neste contexto, surge a necessidade de contar com metodologias onde mostram resíduos que apresentam divergências entre o aminoácido presente na sequência de referência e nas demais sequências utilizadas nesse estudo de caso e que possuem propriedades semelhantes, não alterando a sua função biológica.

O estudo de caso foi realizado com as sequências proteicas: NP_003365.1 (Homo sapiens), NP_035824.1 (Mus musculus) e NP_001001404.1 (Danio rerio) foi executado da seguinte maneira: primeiramente foram adquiridas essas sequências através do banco de dados Uniprot (https://www.uniprot.org) utilizando o campo de busca a palavra chave. Com as sequências inseridas no Essex, utilizamos o modo de alinhamento múltiplo, todas

as proteínas desse conjunto foram alinhadas entre si e o resultado é apresentado na figura 36.

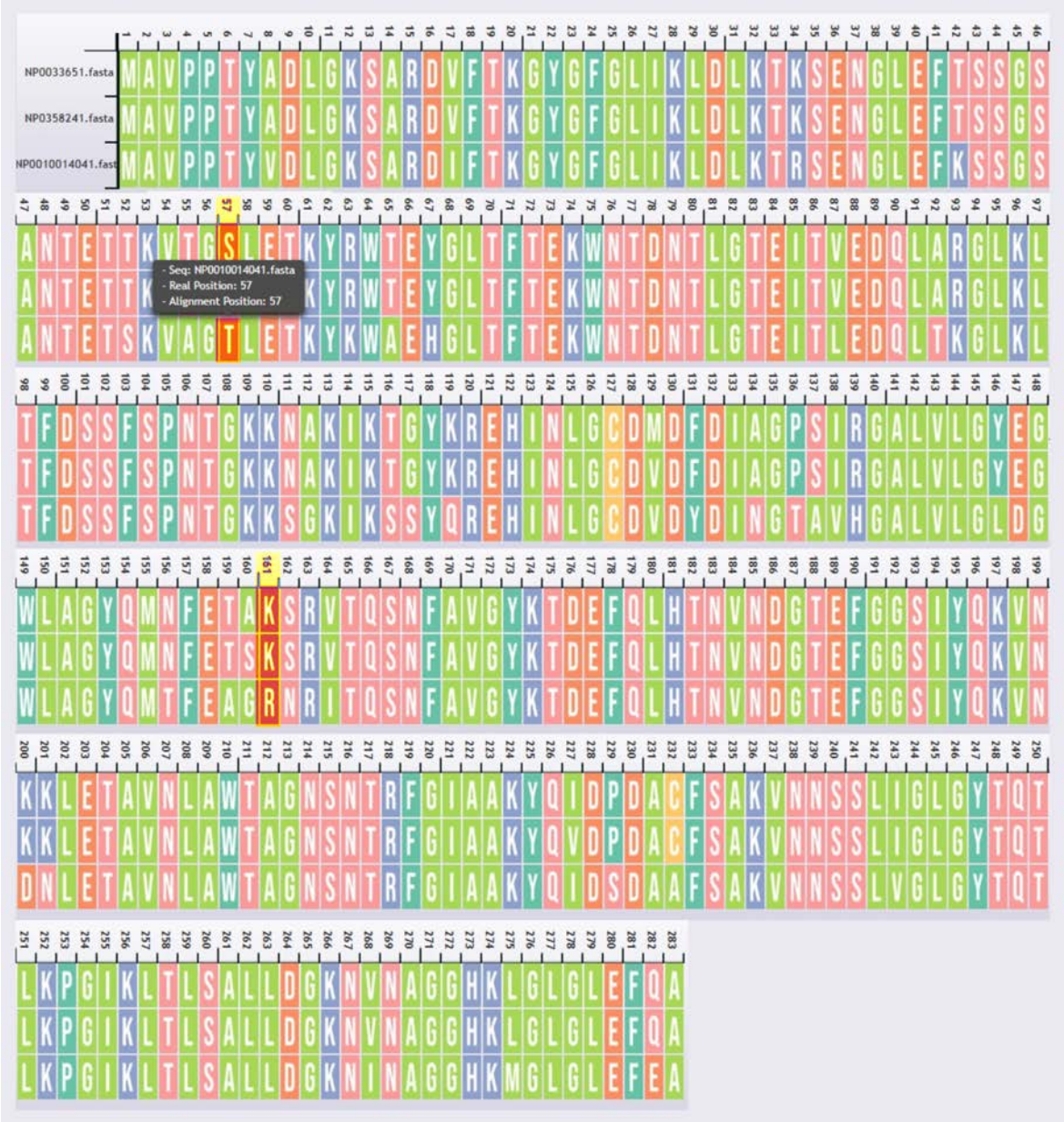

Figura 36: Teste de Caso de Estudo: Alinhamento Múltiplo no Essex

Fonte: Dados do Autor.

Esse mesmo alinhamento é realizado on-line usando os softwares livres Clustal Omega (disponível em https://www.ebi.ac.uk/Tools/msa/clustalo/) e no Promals3D (disponível em http://prodata.swmed.edu/promals3d/promals3d.php).

A figura 37 mostra o resultado realizado na ferramenta Clustal Omega.

Figura 37: Teste de Caso de Estudo: Alinhamento Múltiplo no Clustal Omega

Fonte: Adaptado de https://www.ebi.ac.uk/Tools/msa/clustalo/

Conforme mencionado, o teste também foi realizado na ferramenta Promals3D conforme é apresentado na figura 38.

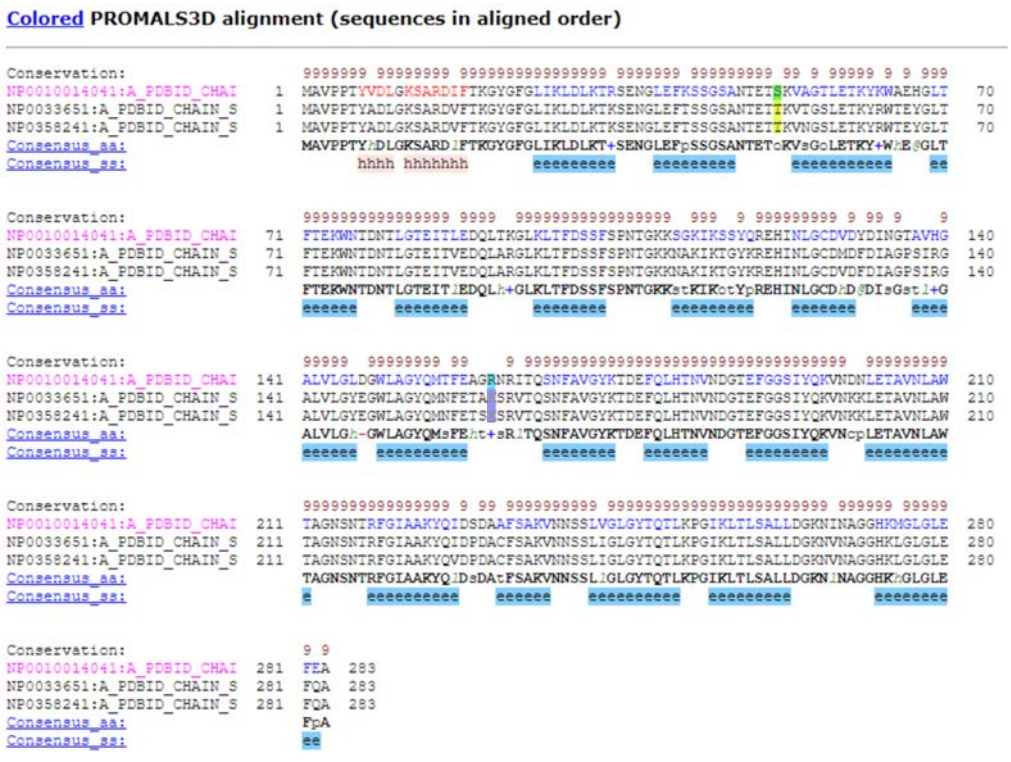

Figura 38: Teste de Caso de Estudo: Alinhamento Múltiplo no Promals3D

Fonte: Adaptado de http://prodata.swmed.edu/promals3d/promals3d.php

No primeiro estudo de caso, foi realizado um alinhamento com três sequências similares do tipo fasta, não havendo ocorrências de inserções de *gaps* tanto de cristalografia quanto de *gaps* de alinhamento que são gerados pela ferramenta para obter um melhor resultado. A ausência dos *gaps* de alinhamento faz com que seja mais fácil de localizar o aminoácido pretendido.

Entretanto, a numeração dos aminoácidos nas ferramentas relacionadas se encontram no lado direito e não em cima de cada aminoácido da sequência, forçando que o usuário localize de maneira manual.

Neste estudo de caso, os pesquisadores tem interesse em verificar a posição dos aminoácidos lisina (K) situado na posição 161 e serina (S) na posição 57.

Foi possível constatar nos testes realizados pelo Essex, Clustal Omega e Promals3D, que o resíduo lisina (K) na posição 161 foi substituído pelo resíduo arginina (R) e o resíduo serina (S) que fica situada na posição 57 foi substituída pelo resíduo tirosina (T) com propriedades semelhantes ou iguais.

5.2 Estudo de Caso 2

Nesse segundo estudo, é realizado alinhamento entre 7 espécies de animais cuja a similaridade é altas, analisando a conservação da proteína mitocondrial ANT1 e se a toxicidade dos nanotubos seria diferente entre as espécies.

O estudo foi feito com as seguintes sequências: Touro, (touro) Bos taurus (NP_777083.1), Humano, Homo sapiens (NP_001142.2), Camundongo, Mus musculus (NP_031476.3), Rato, Rattus norvegicus (NP_445967.1), peixe (zebrafish) Danio rerio (NP_999867.1), Copepodo Lepeophtheirus salmonis (ACO12396.1) e camarão Litopenaeus vannamei (AEZ68611.1) e foram obtidos a partir da base de dados do GenBank (http://www.ncbi.nlm.nih.gov/genbank). Inserimos essas sequências no Essex e utilizamos o método de alinhamento múltiplo, esse conjunto de sequências foram alinhadas entre si conforme é mostrado na figura 39

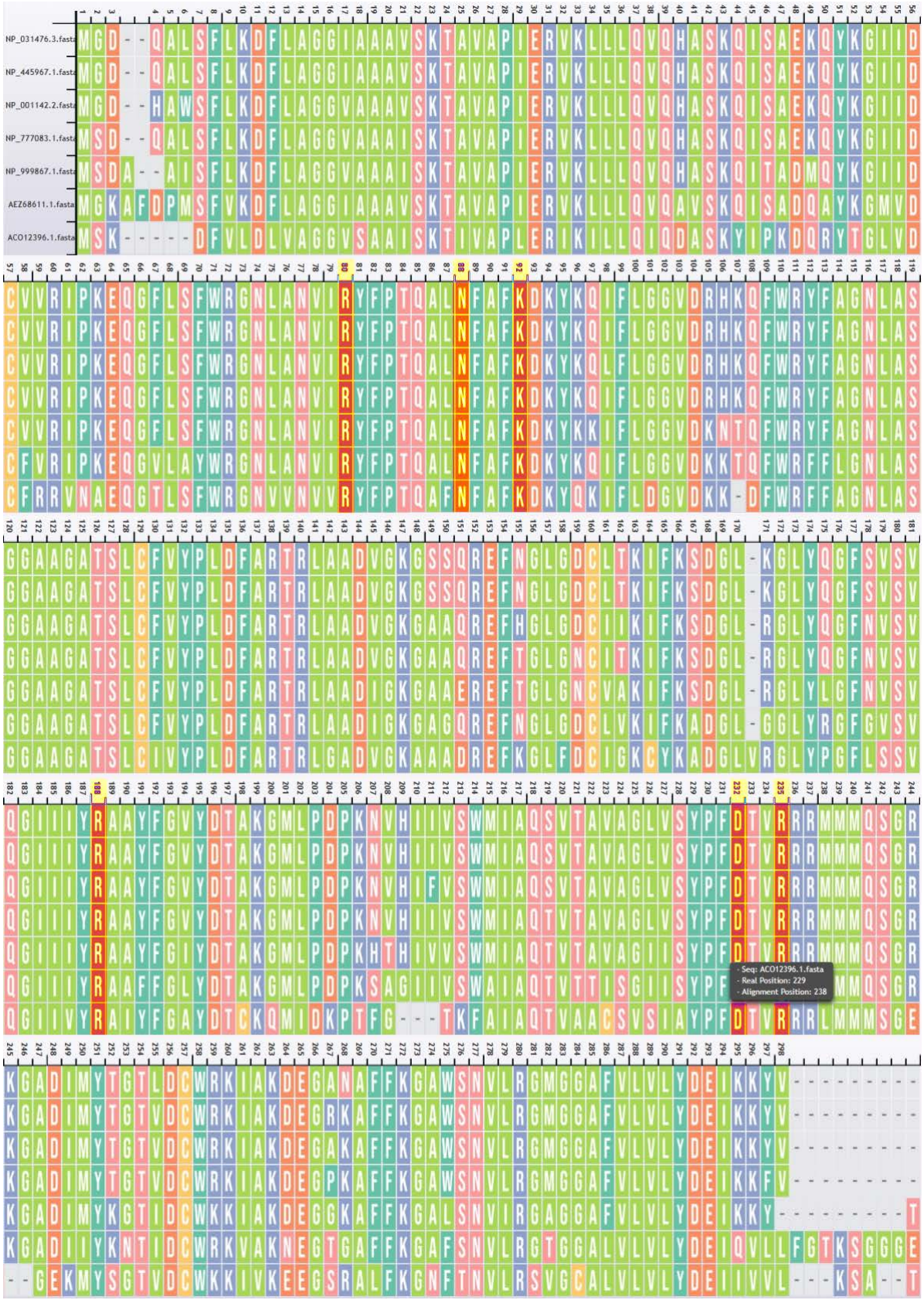

Figura 39: Teste de Caso de Estudo 2: Alinhamento Múltiplo no Essex

Fonte: Dados do Autor.

Esse mesmo alinhamento foi realizado por (GONZÁLEZ-DURRUTHY et al., 2016) utilizando a ferramenta on-line Clustal Omega (disponível em https://www.ebi.ac.uk/Tools/msa/clustalo).

Na figura 40 mostra o alinhamento realizado pela ferramenta Clustal Omega.

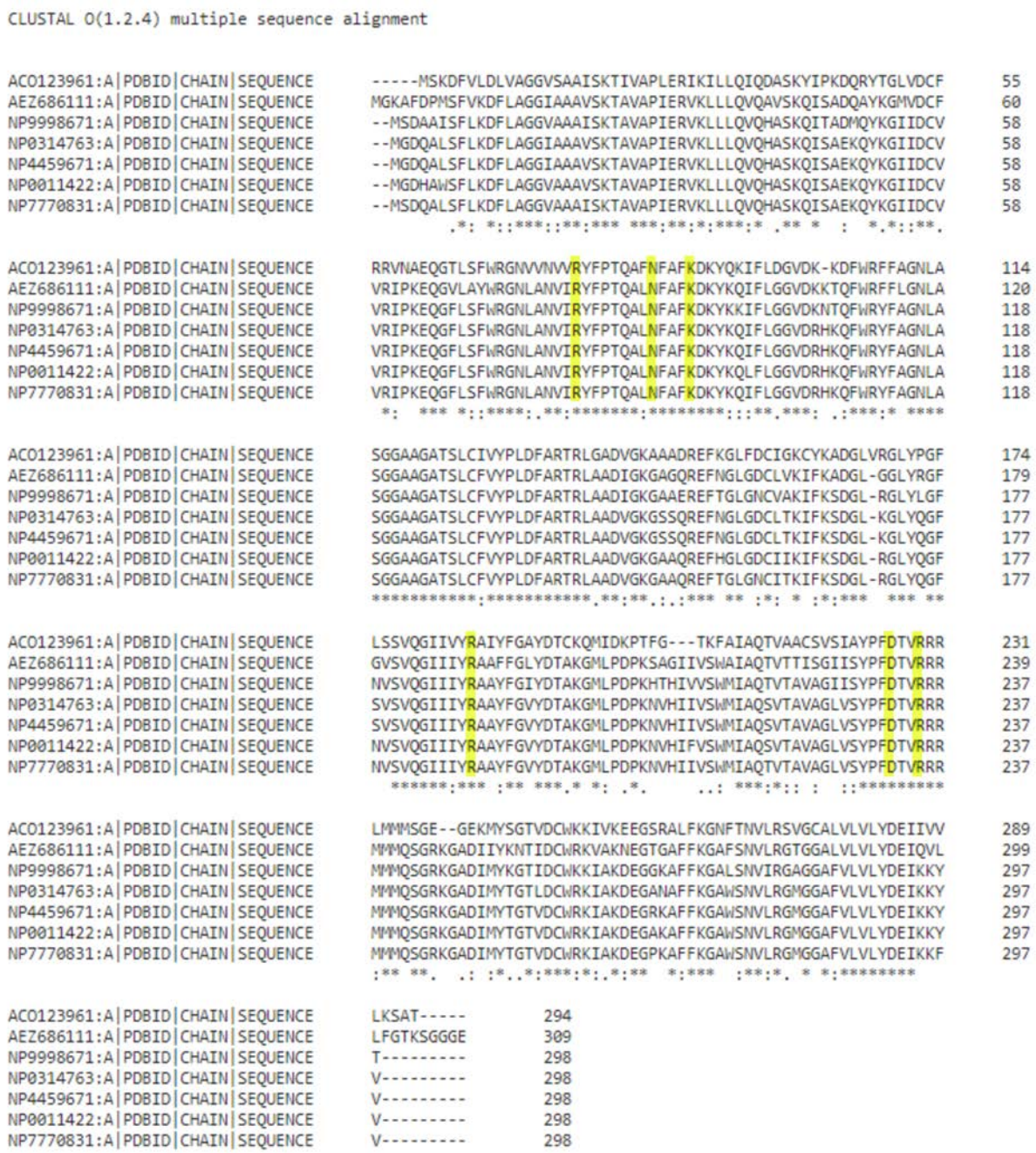

Figura 40: Teste de Caso de Estudo 2: Alinhamento Múltiplo no Clustal Omega

Fonte: Adaptado de https://www.ebi.ac.uk/Tools/msa/clustalo/

O teste foi realizado na ferramenta Promals3D conforme é mostrado na figura 41.

Figura 41: Teste de Caso de Estudo: Alinhamento Múltiplo no Promals3D

Fonte: Adaptado de http://prodata.swmed.edu/promals3d/promals3d.php

Para o alinhamento realizado no segundo estudo de caso, foram introduzidos no Essex, Promals3D e no Clustal Omega cinco sequências, todas as sequências inseridas nesse estudo são do tipo fasta.

Neste estudo de caso, mostra os resíduos do sítio de ligação ativo: Arginina (R) na posição 15, Arginina (R) na posição 163, Arginina (R) na posição 218, Lisina (K) na

posição 12, Lisina (K) na posição 20, Lisina (K) na posição 109, Lisina (K) na posição 113, Lisina (K) na posição 115, Lisina (K) na posição 161, Lisina (K) na posição 174, Lisina (K) na posição 256.

É possível observar que a partir dos resultados gerados pelas ferramentas, não foi encontrada diferença nos resíduos utilizado para esse estudo e constatado ser totalmente conservado em todas as espécies diferentes.

5.3 Estudo de Caso 3

O terceiro estudo de caso é realizado um alinhamento entre 23 sequências de origens distintas para averiguar se existem similaridades entre elas e posteriormente achar resíduos em comum que sejam eficientes na sintese de glicoses para a produção de biocombustível e também validar o alinhamento gerado pelo Essex.

De acordo com MARIANO et al. (2017), β-glicosidases são enzimas que tem um papel importante na degradação de celulose, por essa razão, são consideradas essenciais para a indústria de biocombustíveis. Algumas β-glucosidases são inibidas por glicose, o que torna lento, menos produtiva a ação da enzima, deste modo, β-glucosidases que são mais tolerantes a glicose, são adotadas como alvos para melhorar a produção de biocombustíveis de segunda geração.

Para isso, foi realizado o alinhamento de 23 sequências biológicas utilizando a ferramenta Clustal Omega (SIEVERS et al., 2011) e posteriormente foram escolhidos 22 resíduos conservados de celobiose, encontrando 6 resíduos que são conservados dentre todas as β-glucosidases tolerantes a glicose (H121, N166, E167, Y299, E355 e W402). Foi feito um banco de dados com 5 arquivos estruturais tridimensionais, 23 sequências de β-glucosidases glicose tolerantes, além disso foi modulado 18 sequências de β-glucosidases glicose tolerantes. No intuito que os estudos realizados possam ajudar a desenvolver os biocombustíveis de segunda geração.

Para verificar isso, foi realizado alinhamentos entre os resíduos e detectamos uma subseqüência de consenso GTBGL composta de 22 aminoácidos mais conservados próximos à provável região de ligação da celobiose: HWNEWCLHNLTANYYT-

NEWEWF (H121,W122, N166, E167, W169, C170, L174, H181, N224, L225, T226, A244, N297, Y298, Y299, T300, S302, E355, W402, E409, W410 e F418). Seis resíduos foram conservados em todas as -glicosidases tolerantes à glicose: H121, N166, E167, Y299, E355 e W402.

Na figura 42, apresentamos o alinhamento múltiplo realizado pelo software Clustal Omega (disponível em https://www.ebi.ac.uk/Tools/msa/clustalo), onde resíduos foram conservados em todas as -glicosidases tolerantes à glicose: H121, N166, E167.

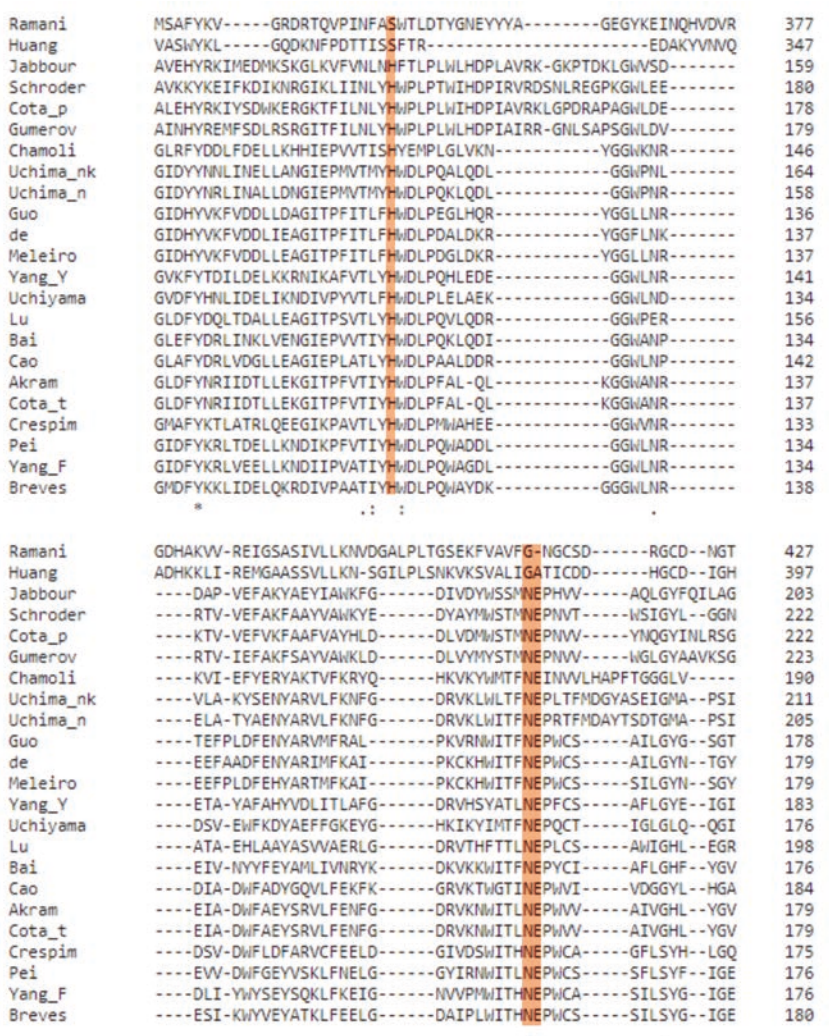

Figura 42: Teste de Caso de Estudo 3: Alinhamento Múltiplo no Clustal Omega - Resíduos: H121, N166, E167

Fonte: Adaptado de https://www.ebi.ac.uk/Tools/msa/clustalo

Na figura 43, mostramos os resíduos conservados em todas as -glicosidases tolerantes

à glicose: Y299, E355 e W402.

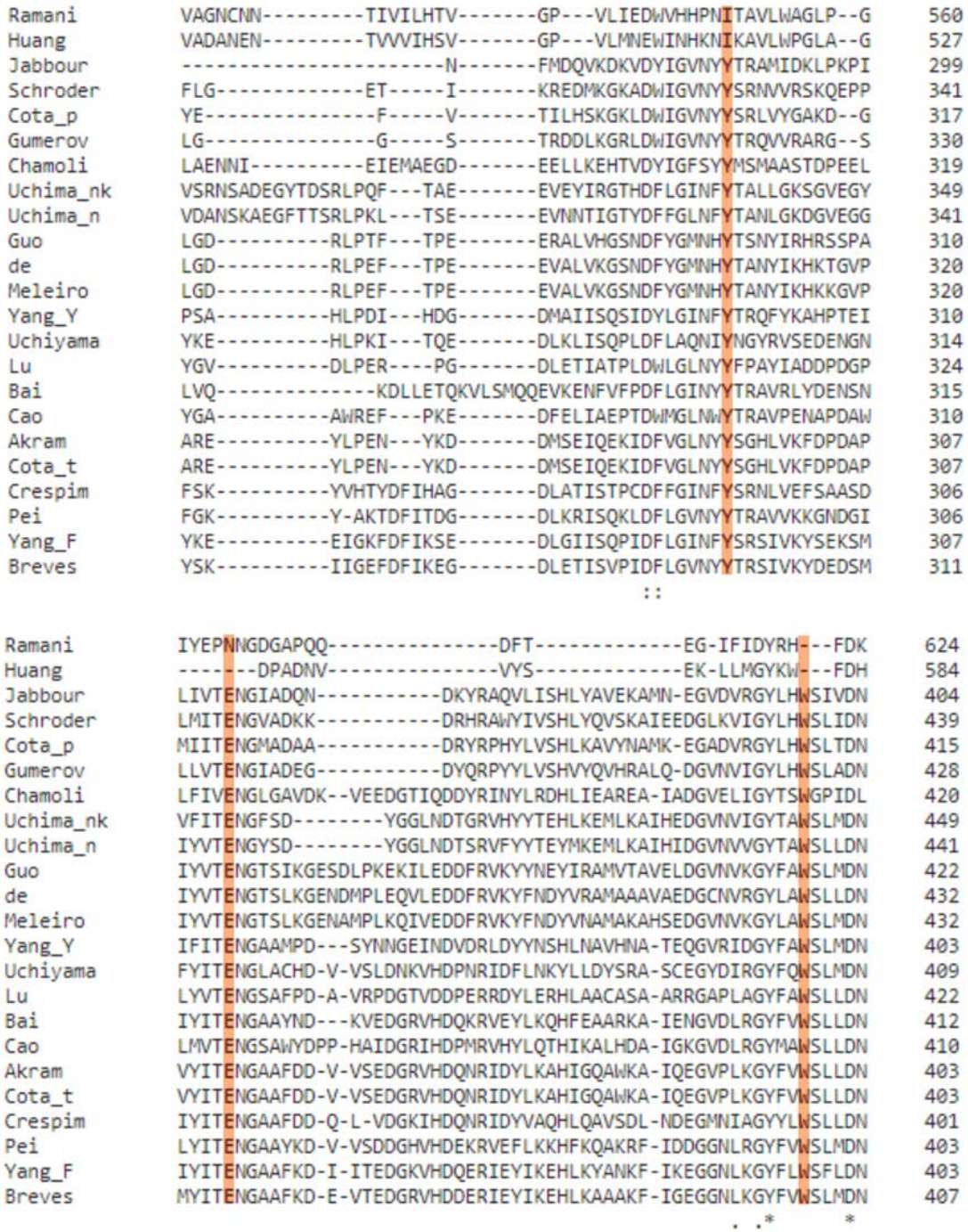

Figura 43: Teste de Caso de Estudo 3: Alinhamento Múltiplo no Clustal Omega - Resíduos: Y299, E355 e W402

Fonte: Adaptado de https://www.ebi.ac.uk/Tools/msa/clustalo

As sequências estão disponíveis nos bancos de dados UniProt (http://www.uniprot.org/) e GenBank (http://www.ncbi.nlm.nih.gov/genbank/) e os

arquivos de estruturas foram obtidos no Banco de Dados de Proteína (BERMAN et al., 2006). Estruturas tridimensionais de proteínas são cruciais para inferir sobre o mecanismo de tolerância à glicose em -glicosidases. No entanto, poucas estruturas experimentais de tolerantes à glicose foram encontradas. Essas sequências e dados estruturais também foram organizados em um banco de dados, chamado betagdb, que estão disponível em: http://bioinfo.dcc.ufmg.br/betagdb.

A partir das bases obtidas no banco de dados, foi realizado um teste utilizando o método de alinhamento múltiplo na ferramenta Essex. Na figura 44, apresentamos o alinhamento realizado pelo Essex, onde resíduos foram conservados em todas as -glicosidases tolerantes à glicose: H121, N166, E167, Y299.

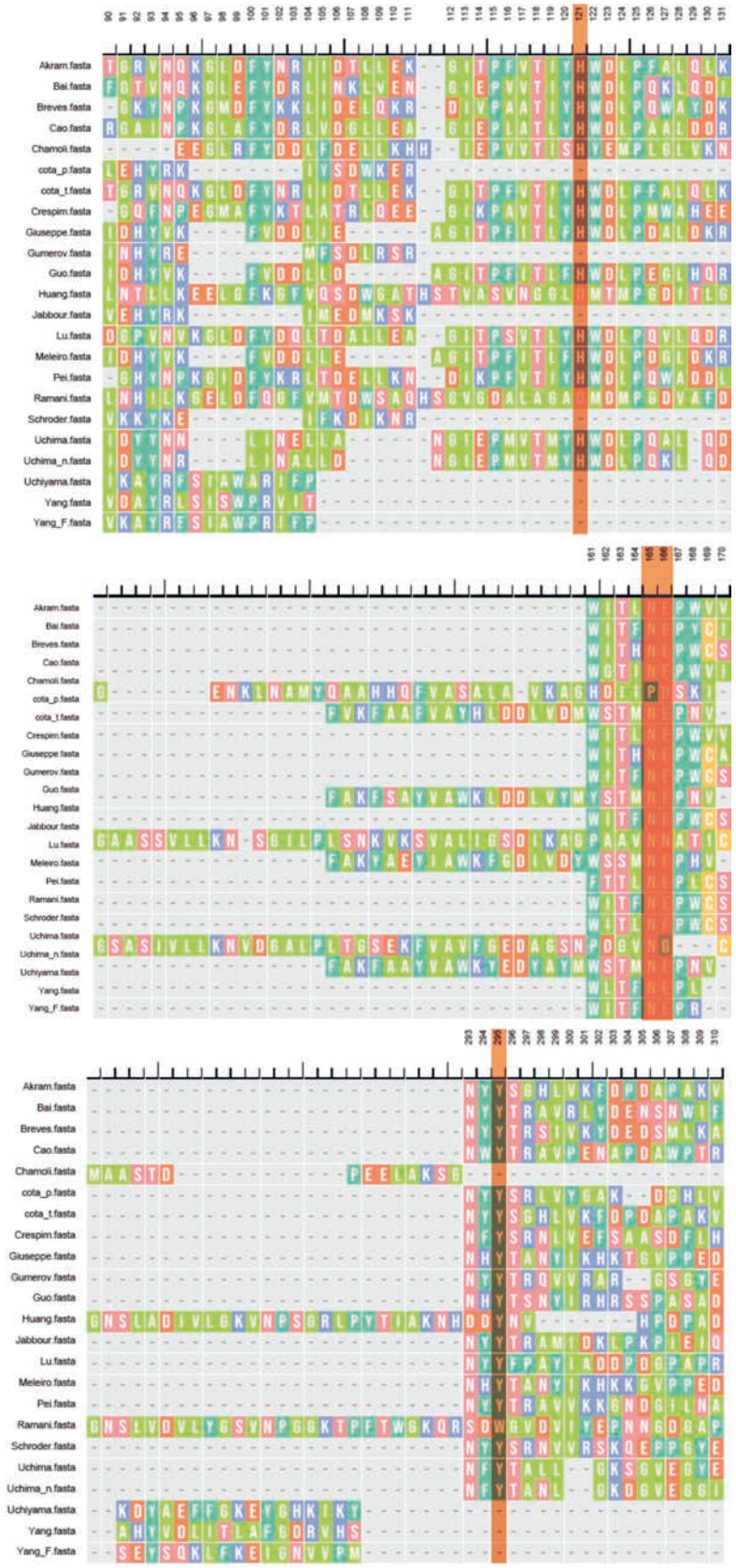

Figura 44: Teste de Caso de Estudo 3: Alinhamento Múltiplo no ESSEX - Resíduos: H121, N166, E167, Y299

Fonte: Dados do Autor.

Na figura 45, mostramos os resíduos conservados em todas as -glicosidases tolerantes à glicose: E355 e W402.

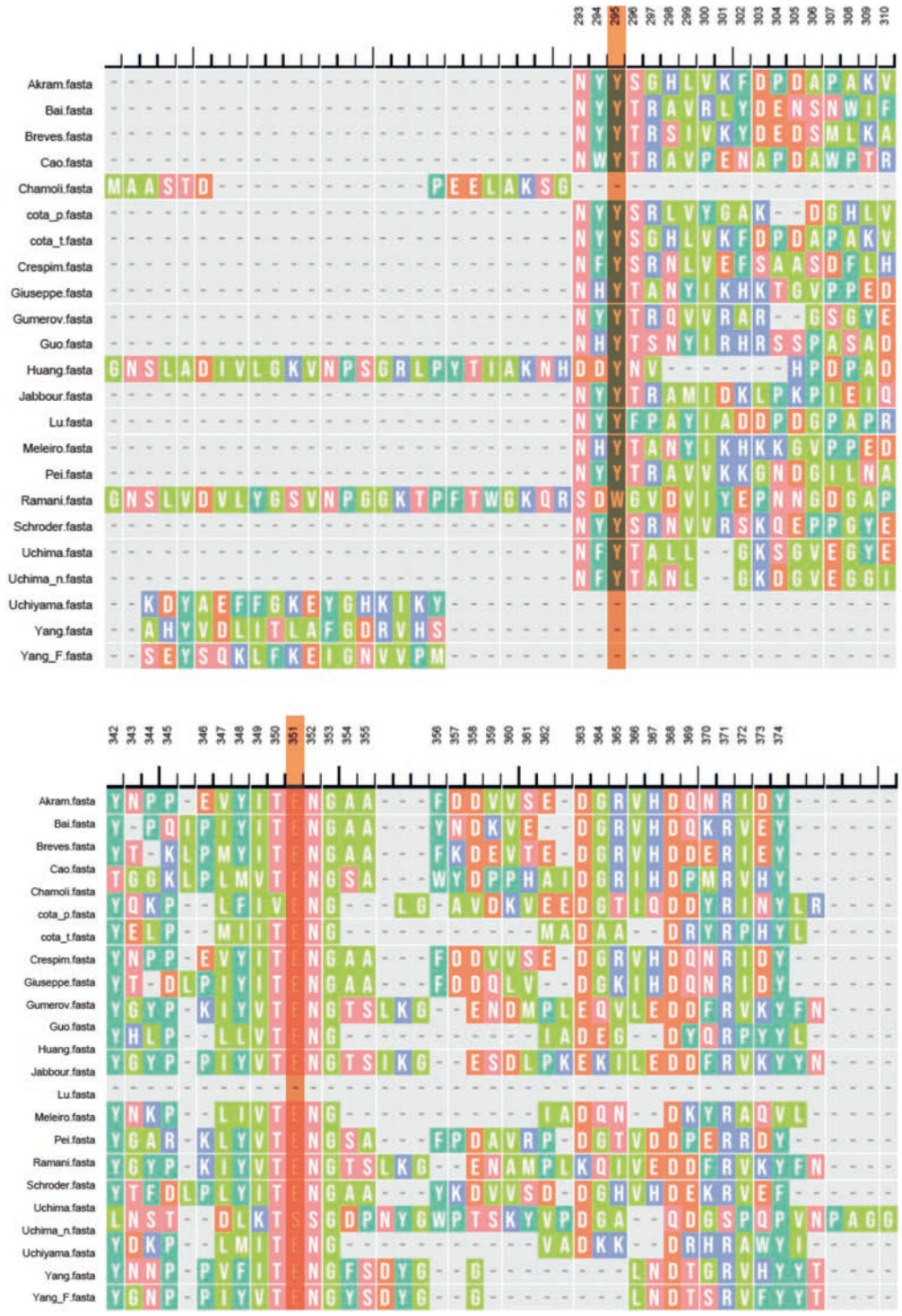

Figura 45: Teste de Caso de Estudo 3: Alinhamento Múltiplo no ESSEX - Resíduos: E355 e W402

Fonte: Dados do Autor.

Através das bases que foram retiradas no banco de dados, também foi gerado um alinhamento múltiplo na ferramenta Promals3D. A figura 46, mostra o alinhamento realizado, onde resíduos foram conservados em todas as -glicosidases tolerantes à glicose: H121, N166, E167, Y299.

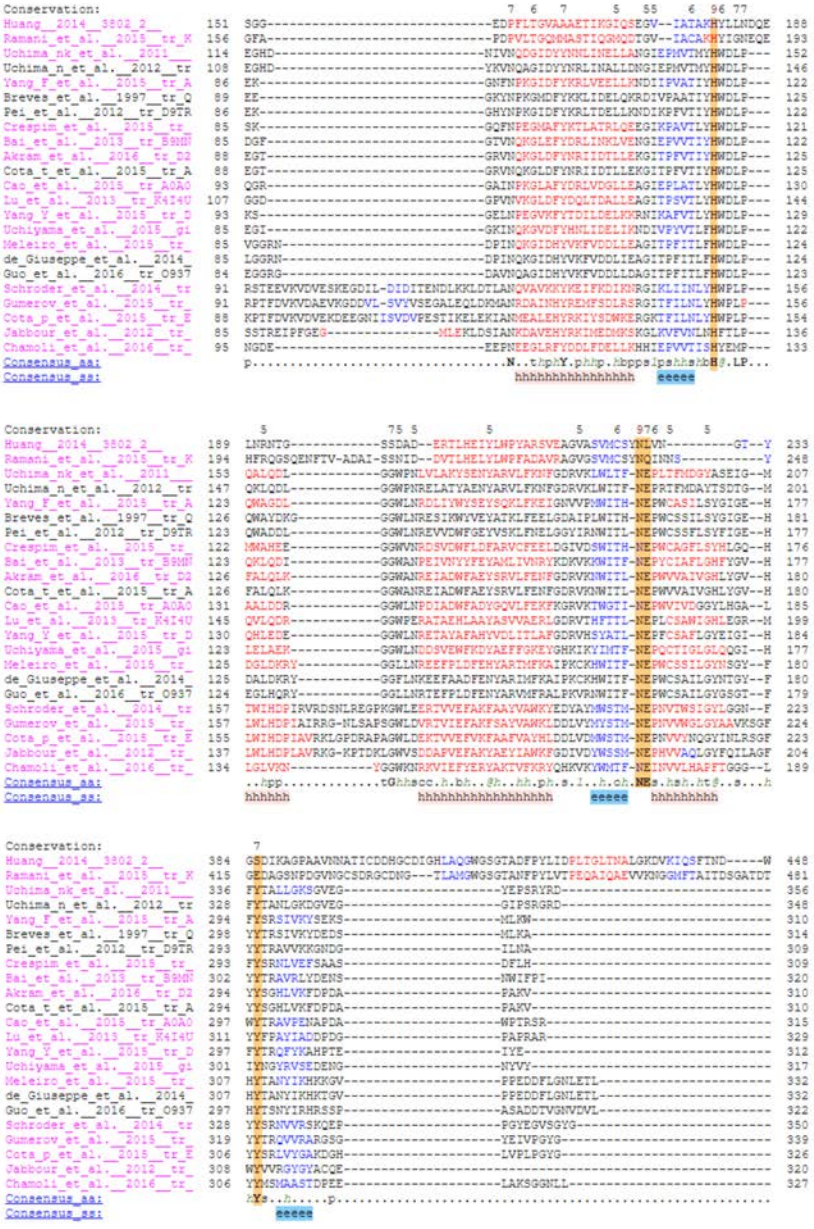

Figura 46: Teste de Caso de Estudo 3: Alinhamento Múltiplo no Promals3D - Resíduos: H121, N166, E167, Y299

Fonte: Adaptado de http://prodata.swmed.edu/promals3d/promals3d.php

Na figura 47, mostra os resíduos conservados em todas as β-glicosidases tolerantes à

glicose: E355 e W402.

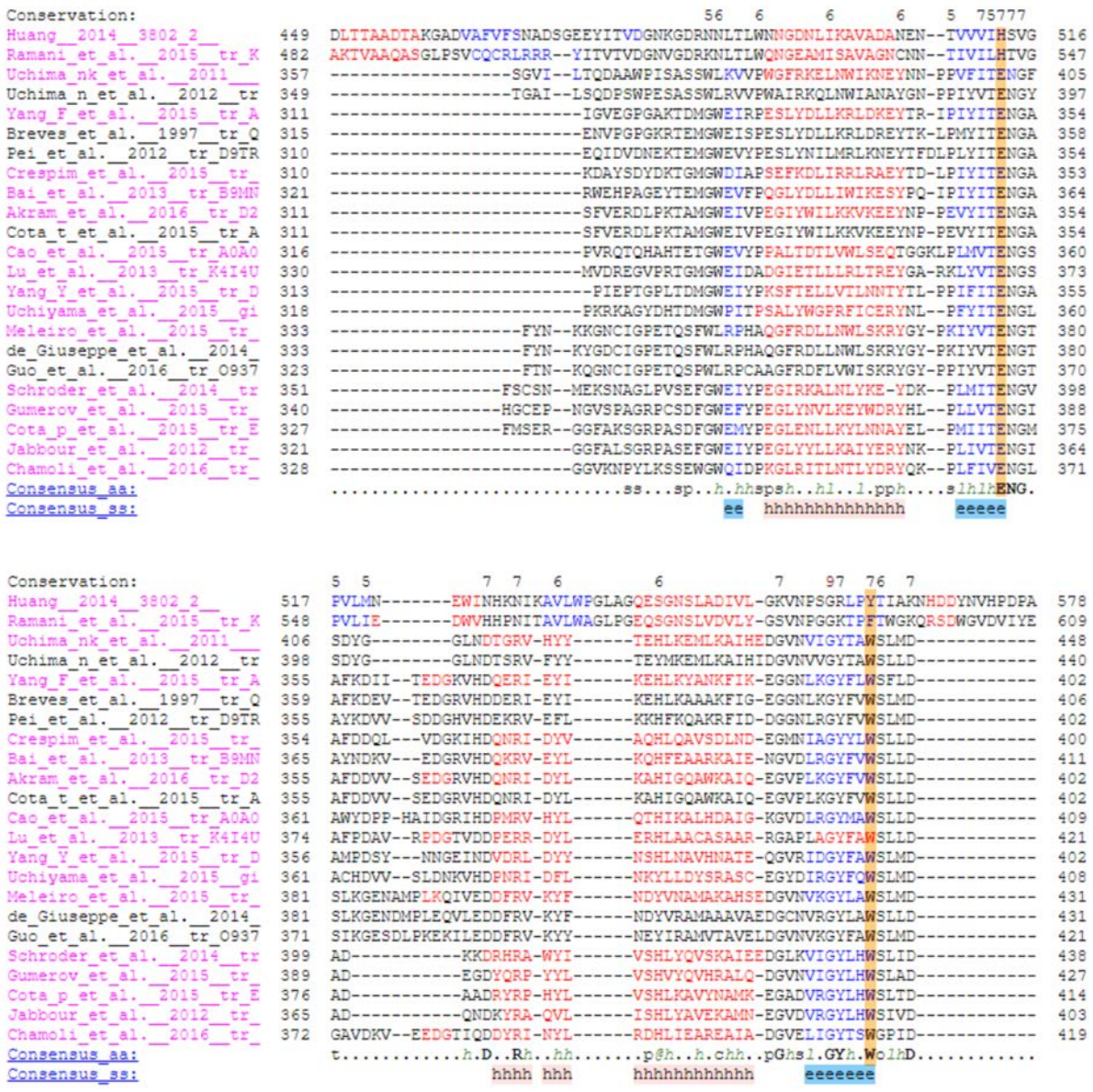

Figura 47: Teste de Caso de Estudo 3: Alinhamento Múltiplo no Promals3D - Resíduos: H121, N166, E167, Y299

Fonte: Adaptado de http://prodata.swmed.edu/promals3d/promals3d.php

Nesse último estudo de caso, foi realizado um teste com 23 sequências no formato PDB para as ferramentas Essex e Promals3D e 23 sequências fasta para o Clustal Omega. O primeiro ponto a ser mencionado é o tempo de execução do alinhamento, o Essex mostrou ser um pouco mais rápido na obtenção do resultado comparado ao Promals3D e Clustal Omega.

Apresentou precisão dos resultados entre as ferramentas e conforme mencionado no estudo de caso anterior, as ferramentas tiveram inserções de *Gaps* de alinhamento e somente a ferramenta Essex apresenta no seu processo de alinhamento os aminoácidos que possuem falhas no processo de cristalografia. Os *gaps* de alinhamento não são contabilizados e os *gaps* de cristalografia são contabilizados, tudo isso para que o usuário possa encontrar de maneira rápida e eficiente a posição correta do aminoácido pretendido.

De um modo geral, as ferramentas Clustal Omega, Promals3D e Essex obtiveram os mesmos resultados de alinhamento. Porém, a numeração das ferramentas Clustal Omega e Promals3D ficam localizadas no lado direito forçando o usuário a contagem de maneira manual do aminoácido pretendido. O Essex conta com um esquema de cores que auxilia a visibilidade de divergências ou convergências no resultado do alinhamento, as ferramentas Clustal Omega e Promals3D não dispõe de um esquema de cores para esse intuito.

Entretanto, no Promals3D identifica aminoácidos que são alfa-hélice e folha-beta respectivamente com as cores vermelho e azul.

Um outro fator importante é que a ferramenta Essex utiliza informações por meio de balões explicativos mostrando a posição real e a posição final após o processo de alinhamento de cada aminoácido e também mostra um resumo da posição real e posição final do aminoácido pretendido em todas sequências que foram inseridas na ferramenta Essex.

Com isso, a ferramenta Essex é a mais eficaz comparada as outras que foram utilizadas nos estudos de casos.

6 CONSIDERAÇÕES FINAIS

Nesse trabalho, mostramos de um modo geral, o conceito de alinhamentos de sequência e os algoritmos de programação dinâmica para obtermos alinhamentos com um ótimo resultado.

O Essex consiste em alinhar e renumerar sequências biológicas no intuito de identificar a posição de um determinado aminoácido de interesse em um processo de alinhamento de sequências biológicas. Essas sequências por sua vez, poderão vir de diversos experimentos e de espécies diferentes.

A comparação de sequências proteicas é uma ferramenta essencial na procura da existência de relações de semelhança entre todo ou parte dessa sequência. Isso é muito comum quando temos uma sequência desconhecida e queremos identificar associando-a um grupo de proteínas de funções conhecidas, comparando essa sequência com outras de um banco de dados, também servem para a predizer as estruturas secundárias de proteínas ou para outras técnicas computacionais como o docking e dinâmica molecular.

6.1 Conclusão e Discussão

Hoje em dia existem várias técnicas e ferramentas para o alinhamento dos mais diversos tipos de sequências de origem biológica. Os algoritmos vão desde de soluções de programação dinâmica, onde o resultado é ótimo, passando por algoritmos híbridos, que combinam soluções ótimas com heurísticas, até soluções completamente heurísticas.

Nesse cenário, geralmente temos interesse específico em alguns aminoácidos, ou nucleotídeos. Como as sequências, de DNA e de estrutura de proteínas, obtidas com di-

ferentes técnicas experimentais, possuem características diferentes, é comum as mesmas apresentarem diferenças no número de elementos, e também nas localizações dos *gaps*. Assim sendo, a tarefa de localizar um elemento de interesse, seja ele um aminoácido ou nucleotídeo, nas diferentes sequências, pode ser bastante complexo.

Apesar de termos disponíveis várias ferramentas de alinhamento, não temos a disposição uma ferramenta que possibilite a renumeração dos elementos da sequência. Essa renumeração poderia facilitar em muito a rápida localização de um elemento de uma sequência de referência em outras sequências alinhadas com a mesma.

Logo, a finalidade desse trabalho é mostrar o funcionamento e exibição da numeração automática de cada aminoácido correspondente na sequência biológica de acordo com a sequência de referência definida pelo usuário que o Essex apresenta. Para retratar as principais vantagens da ferramenta de maneira auxiliar os biólogos que desejam encontrar um aminoácido em um processo de alinhamento de sequências.

O primeiro ponto a ser mencionado nos estudos de caso realizados, teve como um dos objetivos comparar a precisão dos resultados entre as três ferramentas que realizam os alinhamentos múltiplos: Essex, Promals3D, Clustal Omega e nos testes realizados mostraram precisão nos resultados de alinhamentos.

É importante destacar diversas vantagens da utilização do Essex comparada com as demais ferramentas, dentre elas, é possível mencionar a apresentação da régua horizontal enumerando cada aminoácido da sequência com a sequência que foi determinada como referência, após o processo de alinhamento. Essa régua por sua vez, é dinâmica e se adapta sempre de acordo com a sequência que é definida como referência, auxiliando o usuário final a localizar de maneira rápida e eficaz o aminoácido de interesse.

Nesse mesmo contexto, o Essex utiliza de informações representadas por meios de balões explicativos *(tooltips)*, essas informações consistem em posição real que seria a posição original que o aminoácido se encontra na sequência e posição de alinhamento que refere-se sobre a posição que o aminoácido se encontra logo após o processo de alinhamento. A utilização do esquema de cores que a ferramenta disponibiliza, ajuda o usuário na visibilidade de divergências ou convergências no resultado de alinhamento.

Um fator importante é que a ferramenta Essex conta com uma série de opções para auxiliar o usuário a exportar os seus alinhamentos. É possível armazenar essa informação em 3 formatos diferentes: TXT, HTML e PDF. Além disso, também disponibiliza a função *Popups* que é baseado em um resumo da posição real e final do aminoácido pretendido em todas sequências inseridas no alinhamento.

Dentre as inúmeras vantagens que pode obter na utilização da ferramenta, a principal dela é que além de apresentar as inserções de *Gaps* de alinhamento, o Essex mostra em seu processo de alinhamento os aminoácidos que tiveram falhas no processo de cristalografia, o que torna uma funcionalidade exclusiva da ferramenta. Os *gaps* de alinhamento não são contabilizados, porém os *gaps* de cristalografia são contabilizados, isso tudo é oferecido pela ferramenta Essex para que os biólogos não façam esse procedimento de maneira manual, pois há grande chance de não obter o resultado com sucesso na busca da localização do aminoácido escolhido.

7 TUTORIAL

7.1 Entrada

A Figura 48 apresenta a tela inicial onde o usuário irá submeter as sequências para que a ferramenta realize o alinhamento.

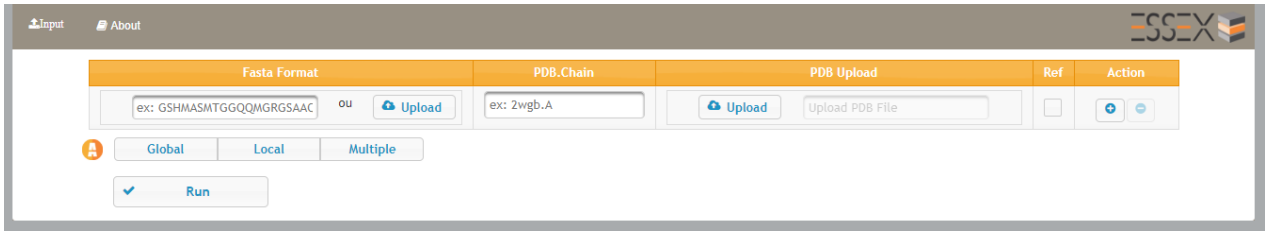

Figura 48: Tela de Input do Essex

Fonte: Dados do Autor.

O Essex aceita os seguintes formatos.

- **Fasta Format:** O usuário poderá inserir a sequência em ”Fasta Format”ou fazendo o upload do mesmo.
- **PDB/Chain:** Faz busca no Banco de Dados do RCSB Protein Data Bank a partir do nome da proteína e sua respectiva família para realizar o alinhamento.
- **Upload:** Usuário faz o upload do arquivo de estrutura no formato PDB e posteriormente escolherá a informação a ser extraída do PDB (Atom ou SeqRes).
- **Ref:** Usuário determina a sequência referência, cada sequência é alinhada com a referência.

- **Action:** Para adicionar ou remover sequências da tabela dinâmica.
- **Alinhamento:** Determina qual método de alinhamento será utilizado.

7.2 Saída

A saída é dada a partir do método de alinhamento que o usuário optou.

- **Alinhamento Global:** Caso usuário tenha escolhido mais de uma sequência e determinando uma delas como referência, o Essex dividirá em abas os respectivos alinhamentos para melhor entendimento. É exibido na figura 49, um exemplo de alinhamento global das sequências de código PDB: 3VIJ, 3HIV e 1HIV.

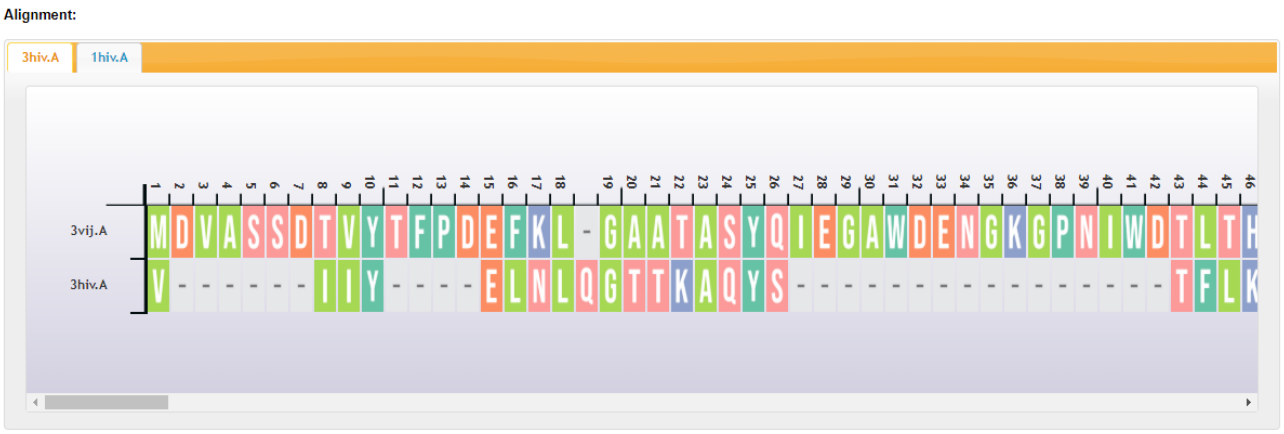

Figura 49: Exemplo de alinhamento múltiplo das proteínas de cód. PDB: 3vij, 3hiv e 1hiv

Fonte: Dados do Autor.

Exemplo: Seq1, Seq2(REF), Seq3, Seq4

As abas serão organizadas da seguinte forma: Seq2(REF) x Seq1, Seq2(REF) x Seq3 e Seq2(REF) x Seq4. Conforme é exibido na figura 50:

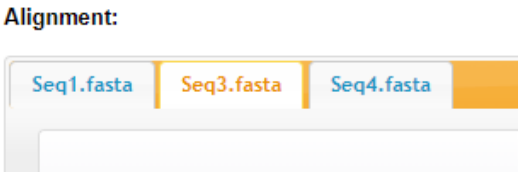

Figura 50: Divisão de abas

Fonte: Dados do Autor.

- **Alinhamento Local:** A figura 51 mostra que o processo é bem parecido com o alinhamento global, o que diferencia é que o programa apresenta somente a região que obteve o maior score de alinhamento.

  Uma outra particularidade no processo de alinhamento local é referente a régua horizontal que enumera cada aminoácido, desta vez enumerando somente a região que obteve o maior score, de maneira que usuário final saiba a quantidade de aminoácidos que foram extraídos entre as sequências.

Figura 51: Exemplo de alinhamento local das proteínas de código PDB: 3via x 3vij

Fonte: Dados do Autor.

- **Alinhamento Múltiplo:** no método de alinhamento múltiplo é utilizado o algoritmo ClustalW que serve para o alinhamento de três ou mais sequências e é amplamente utilizado pois é computacionalmente eficiente e precisa, produzindo alinhamentos de sequências múltiplas biologicamente significativos de sequências divergentes. Um exemplo de alinhamento múltiplo é exibido na figura 52, onde alinha três sequências de código PDB: 3VIJ, 3HIV e 1HIV.

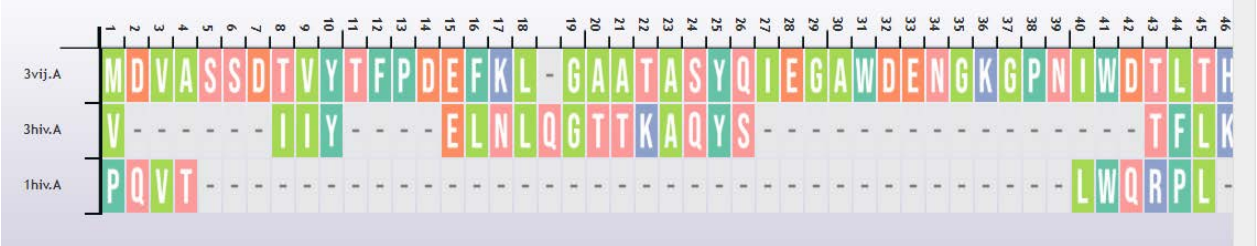

Figura 52: Exemplo de alinhamento múltiplo das proteínas de código PDB: 3vij, 3hiv e 1hiv

Fonte: Dados do Autor.

- **Botões:** Conforme é mencionado na figura 53, o Essex permite que o usuário através de botões, exporte seus alinhamentos nos seguintes formatos: TXT, HTML e PDF Os respectivos botões *Download TXT*, *Download HTML* e *Download PDF* permite ao usuário exportar seus resultados e o botão *Go Back* fará com o que o usuário retorne a tela inicial da ferramenta.

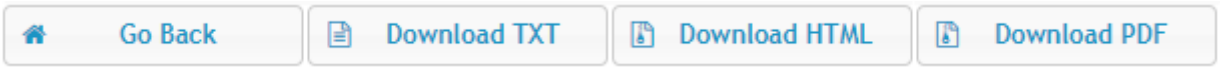

Figura 53: Botões do Essex

Fonte: Dados do Autor.

- **Legenda:** Na figura 54, é possível observar que a ferramenta disponibiliza de um botão *toggle* intitulado como *Subtitles*. Esse botão tem como função de mostrar e esconder a informação de legenda que informa a propriedade de cada aminoácido.

Figura 54: Legendas da ferramenta ESSEX

Fonte: Dados do Autor.

- **Tooltips:** Para auxiliar o usuário da ferramenta a localizar o aminoácido pretendido em um processo de alinhamento, contamos com os *Tooltips*, que consiste em um

balão informativo que permite ao usuário saber a posição real e a posição que ficou após o processo de alinhamento. Para ver essa informação o usuário deverá passar o mouse no aminoácido, como é exibido na figura 55

Figura 55: Balões informativos da ferramenta ESSEX

Fonte: Dados do Autor.

- **Popups:** No mesmo intuito de auxiliar o usuário, o Essex disponibiliza da função *Popups*, que baseia em um resumo da posição real e posição final do aminoácido pretendido em todas sequências que foram inseridas na ferramenta Essex. Para obter essa informação o usuário deverá clicar no aminoácido, como mostra na figura 56.

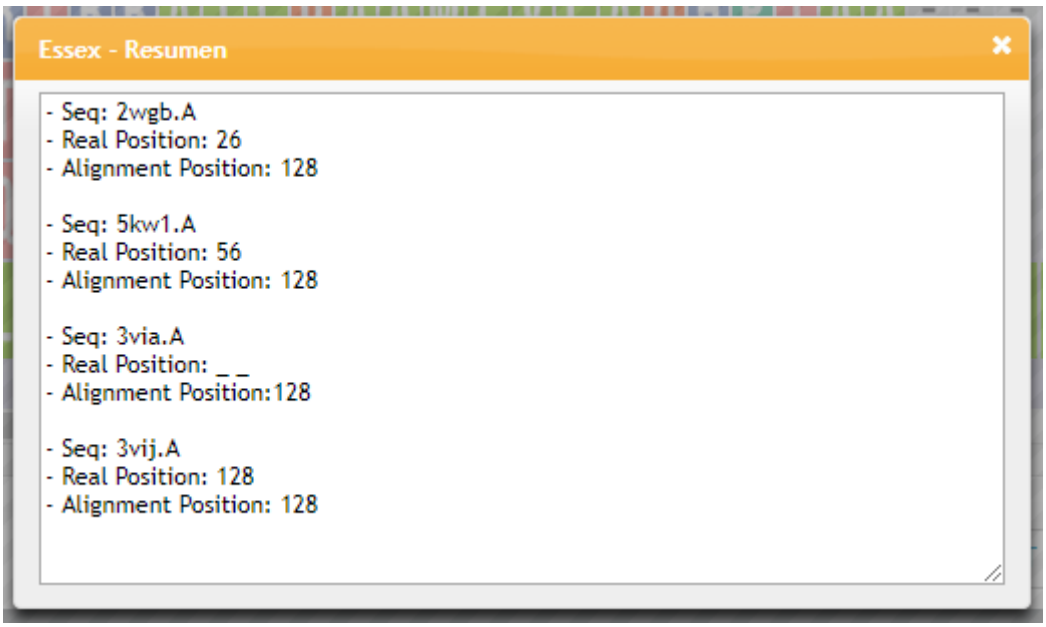

Figura 56: Popup's da ferramenta ESSEX

Fonte: Dados do Autor.

REFERÊNCIAS

ABAGYAN, R.; ORRY, A.; TOTROV, M.; RAUSH, E.; SKORODUMOV, K. ICM User's Guide. **MolSoft, LLC, La Jolla**, [S.l.], 2004.

ACID, S. A. Lehninger principles of biochemistry. , [S.l.], 2004.

AIYAR, A. The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment. In: **Bioinformatics methods and protocols**. [S.l.]: Springer, 2000. p.221–241.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of molecular biology**, [S.l.], v.215, n.3, p.403–410, 1990.

BALDI, P.; BRUNAK, S. **Bioinformatics**: the machine learning approach. [S.l.]: MIT press, 2001.

BERGER, B.; ROZENER, M. **Introduction to computational molecular biology**. [S.l.]: MIT Comp, Biology Ed, 1998.

BERMAN, H.; HENRICK, K.; NAKAMURA, H.; MARKLEY, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. **Nucleic acids research**, [S.l.], v.35, n.suppl_1, p.D301–D303, 2006.

BERMAN, H.; HENRICK, K.; NAKAMURA, H.; MARKLEY, J. L. worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. **Nucleic acids research**, [S.l.], 2007.

BERMAN, H. M.; HENRICK, K.; NAKAMURA, H.; MARKLEY, J. L. The worldwide protein data bank. **Structural bioinformatics. Wiley, John**, [S.l.], p.293–303, 2009.

BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T. N.; WEISSIG, H.; SHINDYALOV, I. N.; BOURNE, P. E. The protein data bank, 1999–. In: **International Tables for Crystallography Volume F**: Crystallography of biological macromolecules. [S.l.]: Springer, 2006. p.675–684.

BERMAN, H.; NAKAMURA, H.; HENRICK, K. The Protein Data Bank (PDB) and the Worldwide PDB http://www. wwpdb. org. **Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics**, [S.l.], 2010.

BRANDEN, C. I. et al. **Introduction to protein structure**. [S.l.]: Garland Science, 1999.

BRANDT, B. W.; HERINGA, J.; LEUNISSEN, J. A. SEQATOMS: a web tool for identifying missing regions in PDB in sequence context. **Nucleic acids research**, [S.l.], v.36, n.suppl_2, p.W255–W259, 2008.

BRITO, R. T. Complexidade de Alinhamento de Seqüências Biológicas. **Trends in Applied and Computational Mathematics**, [S.l.], v.8, n.3, p.319–328, 2007.

BRITO, R. T. de. **Alinhamento de seqüências biológicas**. 2003. Tese (Doutorado em Ciência da Computação) — Master's thesis, Universidade de Sao Paulo.

CARLOS, J.; MEIDANIS. **Introduction to computational molecular biology**. [S.l.]: PWS Pub., 1997. n.04; QH506, S4.

CHENNA, R.; SUGAWARA, H.; KOIKE, T.; LOPEZ, R.; GIBSON, T. J.; HIGGINS, D. G.; THOMPSON, J. D. Multiple sequence alignment with the Clustal series of programs. **Nucleic acids research**, [S.l.], v.31, n.13, p.3497–3500, 2003.

FARIA, P. I. G. de. Montagem de regiões gênicas. , [S.l.], 2013.

FASSIO, A. V.; MARTINS, P. M.; GUIMARÃES, S. d. S.; JUNIOR, S. S.; RIBEIRO, V. S.; MELO-MINARDI, R. C. de; SILVEIRA, S. d. A. Vermont: a multi-perspective visual interactive platform for mutational analysis. **BMC bioinformatics**, [S.l.], v.18, n.10, p.403, 2017.

FENG, D.-F.; DOOLITTLE, R. F. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. **Journal of molecular evolution**, [S.l.], v.25, n.4, p.351–360, 1987.

FERRIER, D.; HARVEY, R. Lippincotts illustrated review biochemistry. **Walters Kluwer: Lippincott Williams and Wilkins**, [S.l.], 2011.

FITCH, W. M. Distinguishing homologous from analogous proteins. **Systematic zoology**, [S.l.], v.19, n.2, p.99–113, 1970.

FITCH, W. M. Homology: a personal view on some of the problems. **Trends in genetics**, [S.l.], v.16, n.5, p.227–231, 2000.

FLANAGAN, D. **JavaScript**: the definitive guide. [S.l.]: "O'Reilly Media, Inc.", 2006.

GOLLERY, M. Bioinformatics: Sequence and Genome Analysis, David W. Mount. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2004, 692 pp., $75.00, paperback. ISBN 0-87969-712-1$. **Clinical Chemistry**, $[S.l.], v.51, n.11, p.2219--2219, 2005$.

GONZÁLEZ-DURRUTHY, M.; WERHLI, A. V.; CORNETET, L.; MACHADO, K. S.; GONZÁLEZ-DÍAZ, H.; WASILIESKY, W.; RUAS, C. P.; GELESKY, M. A.; MONSERRAT, J. M. Predicting the binding properties of single walled carbon nanotubes (SWCNT) with an ADP/ATP mitochondrial carrier using molecular docking, chemoinformatics, and nano-QSBR perturbation theory. **RSC Advances**, [S.l.], v.6, n.63, p.58680–58693, 2016.

GONZÁLEZ-DURRUTHY, M.; WERHLI, A. V.; SEUS, V.; MACHADO, K. S.; PAZOS, A.; MUNTEANU, C. R.; GONZÁLEZ-DÍAZ, H.; MONSERRAT, J. M. Decrypting strong and weak single-walled carbon nanotubes interactions with mitochondrial voltage-dependent anion channels using molecular docking and perturbation theory. **Scientific Reports**, [S.l.], v.7, n.1, p.13271, 2017.

HIGGINS, D. G.; THOMPSON, J. D.; GIBSON, T. J. [22] Using CLUSTAL for multiple sequence alignments. In: **Methods in enzymology**. [S.l.]: Elsevier, 1996. v.266, p.383–402.

HOLLAND, R. C.; DOWN, T. A.; POCOCK, M.; PRLIĆ, A.; HUEN, D.; JAMES, K.; FOISY, S.; DRÄGER, A.; YATES, A.; HEUER, M. et al. BioJava: an open-source framework for bioinformatics. **Bioinformatics**, [S.l.], v.24, n.18, p.2096–2097, 2008.

HRUBY, V. J. Chemistry and Biochemistry of the Amino Acids. **Journal of Pharmaceutical Sciences**, [S.l.], v.75, n.3, p.323, 1986.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, [S.l.], v.22, n.12, p.2577–2637, 1983.

LARKIN, M. A.; BLACKSHIELDS, G.; BROWN, N.; CHENNA, R.; MCGETTIGAN, P. A.; MCWILLIAM, H.; VALENTIN, F.; WALLACE, I. M.; WILM, A.; LOPEZ, R. et al. Clustal W and Clustal X version 2.0. **bioinformatics**, [S.l.], v.23, n.21, p.2947–2948, 2007.

LESK, A. M. **Introduction to protein architecture**: the structural biology of proteins. [S.l.]: Oxford University Press Oxford, 2001.

LI, W.; COWLEY, A.; ULUDAG, M.; GUR, T.; MCWILLIAM, H.; SQUIZZATO, S.; PARK, Y. M.; BUSO, N.; LOPEZ, R. The EMBL-EBI bioinformatics web and programmatic tools framework. **Nucleic acids research**, [S.l.], v.43, n.W1, p.W580–W584, 2015.

LIPMAN, D. J.; ALTSCHUL, S. F.; KECECIOGLU, J. D. A tool for multiple sequence alignment. **Proceedings of the National Academy of Sciences**, [S.l.], v.86, n.12, p.4412–4415, 1989.

MARIANO, D.; LEITE, C.; SANTOS, L.; MARINS, L.; MACHADO, K.; WERHLI, A.; LIMA, L.; MELO-MINARDI, R. de. Characterization of glucose-tolerant β-glucosidases used in biofuel production under the bioinformatics perspective: a systematic review. **GENETICS AND MOLECULAR RESEARCH**, [S.l.], v.16, n.3, 2017.

MCREE, D. E. **Practical protein crystallography**. [S.l.]: Elsevier, 1999.

MCWILLIAM, H.; LI, W.; ULUDAG, M.; SQUIZZATO, S.; PARK, Y. M.; BUSO, N.; COWLEY, A. P.; LOPEZ, R. Analysis tool web services from the EMBL-EBI. **Nucleic acids research**, [S.l.], v.41, n.W1, p.W597–W600, 2013.

MULLAN, L. Pairwise sequence alignment—it's all about us! **Briefings in bioinformatics**, [S.l.], p.113–115, 2006.

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**, [S.l.], v.48, n.3, p.443–453, 1970.

OTTO, M.; THORNTON, J. et al. Bootstrap· The world's most popular mobile-first and responsive front-end framework.'. **Getbootstrap. com**, [S.l.], 2015.

PALME, J.; HOPMANN, A.; SHELNESS, N. **MIME Encapsulation of Aggregate Documents, such as HTML (MHTML)**. [S.l.: s.n.], 1999.

PDFBOX, A. **Apache PDFBox**.

PEI, J.; GRISHIN, N. V. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. **Bioinformatics**, [S.l.], v.23, n.7, p.802–808, 2007.

PEI, J.; GRISHIN, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. In: **Multiple Sequence Alignment Methods**. [S.l.]: Springer, 2014. p.263–271.

PEI, J.; KIM, B.-H.; GRISHIN, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. **Nucleic acids research**, [S.l.], v.36, n.7, p.2295–2300, 2008.

POCOCK, M.; DOWN, T.; HUBBARD, T. BioJava: open source components for bioinformatics. **ACM Sigbio Newsletter**, [S.l.], v.20, n.2, p.10–12, 2000.

PROSDOCIMI, F.; COUTINHO, G.; NINNECW, E.; SILVA, A. F.; REIS, A. N. dos; MARTINS, A. C.; SANTOS, A. C. F. dos; JÚNIOR, A. N.; CAMARGO FILHO, F. Bioinformática: manual do usuário. **Biotecnologia Ciência & Desenvolvimento**, [S.l.], v.29, p.12–25, 2002.

RASMOL. **Schematic Color Amino Acid**. [S.l.]: RasMol Colors and Color Schemes, 2002.

ROCHA, E. Folhas de BioInformática e Análise de sequências. , [S.l.], 2011.

SCHULZ, G. E.; SCHIRMER, R. H. **Principles of protein structure**. [S.l.]: Springer Science & Business Media, 2013.

SIEVERS, F.; HIGGINS, D. G. Clustal omega. **Current protocols in bioinformatics**, [S.l.], v.48, n.1, p.3–13, 2014.

SIEVERS, F.; WILM, A.; DINEEN, D.; GIBSON, T. J.; KARPLUS, K.; LI, W.; LOPEZ, R.; MCWILLIAM, H.; REMMERT, M.; SÖDING, J. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. **Molecular systems biology**, [S.l.], v.7, n.1, p.539, 2011.

SILVEIRA, S. A.; FASSIO, A. V.; GONÇALVES-ALMEIDA, V. M.; LIMA, E. B. de; BARCELOS, Y. T.; ABURJAILE, F. F.; RODRIGUES, L. M.; MEIRA JR, W.; MELO-MINARDI, R. C. de. Vermont: Visualizing mutations and their effects on protein physicochemical and topological property conservation. In: BMC PROCEEDINGS, 2014. **Anais...** [S.l.: s.n.], 2014. v.8, n.2, p.S4.

SIPPL, M. J.; WIEDERSTEIN, M. A note on difficult structure alignment problems. **Bioinformatics**, [S.l.], v.24, n.3, p.426–427, 2008.

SMITH, T. F.; WATERMAN, M. S. Comparison of biosequences. **Advances in applied mathematics**, [S.l.], v.2, n.4, p.482–489, 1981.

SONNHAMMER, E. L.; KOONIN, E. V. Orthology, paralogy and proposed classification for paralog subtypes. **TRENDS in Genetics**, [S.l.], v.18, n.12, p.619–620, 2002.

TELLES, G. P.; ALMEIDA, N. F.; MARTINEZ, F. H. V. Algoritmos e heurısticas para comparaçoes exata e aproximada de sequências. **XXIV Jornadas de Atualizaçao em Informática. A. Loureiro and M. Barcelos.(Org.)**, [S.l.], p.1545–1587, 2005.

TICONA, W. G. C. Aplicação de Algoritmos Genéticos Multi-Objetivo para Alinhamento de Seqüências Biológicas. **Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, SP**, [S.l.], 2003.

VENNERS, B. **The java virtual machine**. [S.l.]: McGraw-Hill, New York, 1998.

VERLI, H. Bioinformática: da biologia à flexibilidade molecular. , [S.l.], 2014.

VUKOTIC, A.; GOODWILL, J. **Apache Tomcat 7**. [S.l.]: Springer, 2011.

ZAFALON, G. F. D. Algoritmos de alinhamento múltiplo e técnicas de otimizaçao para esses algoritmos utilizando Ant Colony. , [S.l.], 2009.