RESSALVA

Atendendo solicitação do
autor, o texto completo
desta dissertação será
disponibilizado somente a
partir de 07/08/2021.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

José Augusto Sabino de Oliveira

Melhoramento do resultado final do alinhamento múltiplo de sequências por meio de métodos de rearranjo de árvores e do princípio da evolução mínima

São José do Rio Preto
2019

José Augusto Sabino de Oliveira

Melhoramento do resultado final do alinhamento múltiplo de sequências por meio de métodos de rearranjo de árvores e do princípio da evolução mínima

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Orientador: Prof. Dr. Geraldo Francisco Donegá Zafalon

São José do Rio Preto
2019

O48m

Oliveira, José Augusto Sabino de

Melhoramento do resultado final do alinhamento múltiplo de sequências por meio de métodos de rearranjo de árvores e do princípio da evolução mínima / José Augusto Sabino de Oliveira. -- São José do Rio Preto, 2019

95 f. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Geraldo Francisco Donegá Zafalon

1. Ciência da computação. 2. Bioinformática. 3. Alinhamento de sequências. 4. Rearranjo de Árvore. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

José Augusto Sabino de Oliveira

Melhoramento do resultado final do alinhamento múltiplo de sequências por meio de métodos de rearranjo de árvores e do princípio da evolução mínima

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Comissão Examinadora

Prof. Dr. Geraldo Francisco Donegá Zafalon
UNESP – Câmpus de São José do Rio Preto
Orientador

Prof. Dr. Jorge Rady de Almeida Junior
USP – Câmpus de São Paulo

Prof. Dr. Leandro Alves Neves
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
07 de agosto de 2019

Ao meu pai, que sempre me guiou pelo caminho correto, acreditou e me inspirou na busca pelo conhecimento.
A minha esposa Maria, pela força e companheirismo ao longo dessa caminhada.
Para Alexandre, Leonardo, Mirella e Arthur.

Agradecimentos

À Deus por ser a luz que me guia e não me deixa desanimar frente aos desafios;

À toda minha família, por toda força e confiança;

Em especial, meu pai que sempre me guia e inspira a fazer sempre o melhor;

A minha esposa pelo companheirismo e pelo tempo abdicado;

Aos meus filhos, por serem o motivo de toda minha luta;

Ao meu orientador, por ter acredito e viabilizado a realização desse trabalho;

A todos do Grupo de Computação Cientifica da Unesp de São José do Rio Preto;

Aos colegas do Laboratório de Bioinformática da Unesp de São José do Rio Preto;

A todos os meus professores.

"*Aquele que ousa perder uma hora de seu tempo não sabe o valor da vida.*"
(CHAPIN; CHAPIN, 1935)

Resumo

Os recentes avanços científicos e tecnológicos proporcionaram um grande aumento dos dados biológicos disponíveis para análise e técnicas manuais são inviáveis para executar a análise desses dados e, assim, a Bioinformática surgiu devido a necessidade de analisar-se esse volume de dados com ferramentas computacionais e possibilitar inferências relevantes. Neste contexto, técnicas de alinhamento de sequências são ferramentas imprescindíveis. No entanto, alinhar sequências tem alto custo computacional e neste cenário, adota-se o Alinhamento Múltiplo de Sequências no qual diversas heurísticas são adotadas para amenizar o problema do custo computacional. Uma destas é o Alinhamento Progressivo que funciona basicamente em três fases: alinhamento par a par, construção da árvore guia e o alinhamento de perfis. Diversos estudos destacam a importância da árvore guia e de refiná-la para manter as sequências mais similares em ramificações próximas e assim, produzir um resultado com melhor significância biológica. Assim, no presente trabalho apresenta-se conceitos e técnicas essenciais da Bioinformática e de construção da árvore guia e, por fim apresenta-se um método de refinamento da árvore guia aplicado em ferramentas que usam o Alinhamento Progressivo e que trabalha entre as etapas de construção da árvore guia e do profile final. O método recebe como entrada uma árvore guia e sua matriz de distâncias produzida pelo Alinhamento Progressivo, promove mudanças em suas ramificações internas e avalia se a probabilidade da árvore produzida é mais próxima da árvore verdadeira. Se probabilidade da árvore produzida é maior que a da árvore guia inicial, a nova árvore é devolvida como a árvore guia para a ferramenta de alinhamento construir o profile final, o que proporciona resultados com melhor significado biológico.

Palavras-chave: Bioinformática. Alinhamento Múltiplo de Sequências. Alinhamento Progressivo. Rearranjo de Árvores. Refinamento da Árvore Guia.

Abstract

Recent scientific and technological advances have provided a large increase in the biological data available for analysis and manual techniques are unfeasible to perform the analysis of this data and thus Bioinformatics arose due to the need to analyze this data volume with computational tools and enable relevant inferences. In this context, sequence alignment techniques are essential tools. However, aligning sequences has a high computational cost and in this scenario, Multiple Sequence Alignment is adopted in which several heuristics are adopted to alleviate the computational cost problem. One of these is Progressive Alignment which works basically in three phases: pairwise alignment, guide tree construction and profile alignment. Several studies highlight the importance of the guide tree and refining it to maintain the most similar sequences in close branches and thus produce a result with better biological significance. Thus, in the present work we discuss essential concepts and techniques of Bioinformatics and the construction of the guide tree and, finally, we present a method of refinement of the guide tree applied in tools that use Progressive Alignment and that works between the guide tree build step and the final profile build step. The method receives as input a guide tree and its distance matrix produced by Progressive Alignment, makes changes in its internal branches and evaluates if the probability of the produced tree is closer to the true tree. If the produced tree's probability is greater than that of the initial guide tree, the new tree is returned as the guide tree for the alignment tool to build the final profile, which gives better biological significance results.

Keywords: Bioinformatics. Multiple Sequence Alignment. Progressive Alignment. Tree Rearrangement. Guide Tree Refinement.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AMS *Alinhamento Múltiplo de Sequências*

DNA *Deoxyribonucleic Acid*

FFT *Fast Fourier Transform*

GA *Algoritmos Genéticos*

GPU *Graphics Processing Unity*

GPU *Graphics Processing Unit*

HMM *Hidden Markov Model*

IDE *Integrated Development Environment*

MAFFT *Multiple Alignment using Fast Fourier Transform*

ME *Minimum Evolution Principle*

ML *Maximum-Likelihood*

MP *Maximum-Parsimony*

MUSCLE *Multiple Sequence Comparison by Log-Expectation*

NGS *Next Generation Sequencing*

NJ *Neighbor-Joining*

NNI *Nearest Neighbor Interchange*

PCR *DNA Polimerase*

PD *Programação Dinâmica*

PHYML *Phylogenetic Maximum Likelihood*

Q *Q Score*

RAxML *Randomized Axelerated Maximum Likelihood*

RNA *Ribonucleic Acid*

SA *Simulated Annealing*

SP *Sum-of-Pairs*

SPR *Subtree Pruning and Regrafting*

TBR *Tree Bisection and Reconnection*

TBR *Tree Bisection and Reconnection*

TC *Total Column Score*

TRMLE *Tree Rearrangement + Maximum Likelihood + Minimum Evolution*

UPGMA *Unweighted Pair Group Method using Arithmetic averages*

WSP *Weighted Sum-of-Pairs*

Sumário

Capítulo 1

Introdução

1.1 Contextualização

São muitas as revoluções e descobertas que ocorreram em todas as áreas da ciência no século XX e nesse princípio do século XXI. Entre todas ciências, a Biologia enquadra-se como uma das que mais se destacou, devido aos vários avanços que tem proporcionado (COHEN, 2004). Apesar de haver registros históricos de que os antigos chineses, gregos e egípcios estudavam alguns aspectos dos seres vivos e de estruturas das células em séculos anteriores, a Biologia só teve sua considerável revolução a partir da descoberta da estrutura 3D, em forma de hélice, do DNA, por Watson e Crick em 1953 (WATSON; CRICK et al., 1953).

Por outro lado, a Ciência da Computação é uma área que se desenvolveu no século XX, com importantes contribuições do matemático britânico Alan Turing, que lançou as suas bases ao introduzir, em 1936, um modelo matemático simples, de uma máquina hipotética, para o processo computacional e que atualmente é conhecida como Máquina de Turing (TURING, 1937). Além disso, houve também uma contribuição relevante do matemático húngaro John von Neumann, que trabalhou com seus colegas do Instituto de Estudos Avançados de Princeton, entre 1946 e 1952, para desenvolver um computador que executasse programas armazenados, que foi chamado de IAS

Computer e se tornou o protótipo de todos os computadores modernos (STALLINGS, 2003).

No entanto, a Ciência da Computação não é um fim específico, mas um meio pelo qual as demais ciências podem obter melhores resultados para os seus trabalhos e pesquisas. Dessa forma, solucionando problemas complexos de forma automática, mais rápido do que o trabalho e a mente humana são capazes de realizar. A aplicação das técnicas e ferramentas da Ciência da Computação na área das Ciências Biológicas deu origem a uma área de conhecimento multidisciplinar conhecida como Bioinformática (EDGAR; BATZOGLOU, 2006).

Um dos principais fatores que contribuíram para a evolução da Bioinformática e que a torna cada vez mais importante, relaciona-se com os avanços na medicina. O progresso fundamental na medicina depende da elucidação de como ocorrem os vários processos biológicos, levando-se a uma necessidade crescente e constante de avanços tecnológicos (COHEN, 2004). O projeto Genoma Humano e seus desmembramentos podem ser citados como exemplos de impulsionadores desses avanços citados na Bioinformática (PROSDOCIMI et al., 2002).

Além desses fatores de avanço relatados anteriormente, o surgimento dos sequenciadores de nova geração *NGS (Next Generation Sequencing)* revolucionou ainda mais a pesquisa genômica, devido ao grande volume de dados que ele é capaz de gerar, quando comparado com a tecnologia anterior, a Sanger (LIU et al., 2012). Para se ter um efeito comparativo, o que um NGS produz em um dia, um Sanger produz em anos (BEHJATI; TARPEY, 2013). Esse fato tem aumentando de maneira considerável a quantidade de dados biológicos de sequências de DNA e de proteínas que precisam ser analisados e armazenados. Com esses avanços tecnológicos, tornou-se infactível a análise manual desses dados coletados pelos biólogos (ZAFALON et al., 2015), que têm a necessidade de interpretá-los. Dessa forma, estratégias computacionais para analisá-los tornaram-se essenciais, especialmente considerando-se que os dados obti-

das por meio do sequenciamento são expressados na forma de cadeia de caracteres, ou seja, sequências de símbolos (COHEN, 2004).

No caso do DNA, que é base da hereditariedade, trata-se de um polímero composto por moléculas chamadas nucleotídeos, que podem ser mapeados por quatro bases nitrogenadas: Adenina (A), Citosina (C), Guanina (G) e Timina (T). Uma sequência de DNA é, portanto, especificada completamente por uma sequência composta por um alfabeto de quatro letras A, C, G, T (LIEW; YAN; YANG, 2005). Essa mesma abstração aplica-se aos aminoácidos, no entanto, usa-se um alfabeto de vinte letras, cujas combinações formam as proteínas.

A Bioinformática desempenha um importante papel na coleta e processamento dos dados genômicos obtidos pelos biólogos no estudo das funções proteicas, bem como a determinação das regiões conhecidas como pontos quentes *(hotspots)*, que são relevantes para a indústria farmacêutica (COHEN, 2004). O conhecimento detalhado das estruturas das proteínas facilita o desenvolvimento de medicamentos capazes de atuar nessas regiões de interesse e atacar as causas das potenciais doenças. Desta forma, algumas das tarefas típicas realizadas na Bioinformática incluem indicar a forma e a função de uma proteína a partir de uma determinada sequência de aminoácidos, encontrar todos os genes e proteínas em um dado genoma e determinar os locais na estrutura proteica onde moléculas de drogas podem ser anexadas (COHEN, 2004). Essa percepção apresenta-se como um fator de suma importância em diversos cenários, principalmente no auxílio à eventual cura de diferentes tipos de doenças (KHURI, 2008). Assim, biólogos têm a necessidade do poder computacional para a análise satisfatória dos dados que, geralmente, são obtidos em sua forma bruta e que precisam ser processados para a realização de posteriores inferências (KHURI, 2008).

1.2 Justificativas

Existem vários métodos que são usados em várias ferramentas computacionais para extrair informações genômicas das amostras biológicas sequenciadas e as que mais se destacam são as ferramentas de alinhamento de sequências (EDGAR; BATZOGLOU, 2006). Uma das principais estratégias utilizadas pelos métodos empregados para as análises de sequências biológicas é o Alinhamento Múltiplo de Sequências (AMS) (MORRISON; MORGAN; KELCHNER, 2015). O AMS é uma estratégia útil na previsão de estruturas de proteínas, análise filogenética, predição de função e no desenho dos iniciadores da reação em cadeia da *DNA polimerase (PCR)* (NOTREDAME; HOLM; HIGGINS, 1998).

No entanto, AMS é uma técnica não-determinística e nem sempre garante o resultado exato, mas sim um alinhamento ótimo, ao contrário do que pode ser obtido por meio de técnicas determinísticas, como a Programação Dinâmica (PD). Dentre essas técnicas de PD, destacam-se os algoritmos de Needleman-Wunsch (NEEDLEMAN; WUNSCH, 1970) e de Smith-Waterman (SMITH; WATERMAN, 1981). Estas técnicas são adequadas para tratarem pares de sequências (WANG; JIANG, 1994). Assim, a utilização de estratégias determinísticas para alinhamentos de múltiplas sequências torna-se uma tarefa com alto custo computacional, uma vez que os algoritmos de PD anteriormente citados enquadram-se como problemas da classe dos NP-completos, com complexidade de tempo e espaço $O(L^N)$, em que L é o comprimento da sequência e N é o número de sequências de uma conjunto (WATERMAN; SMITH; BEYER, 1976; WANG; JIANG, 1994; EDGAR, 2004a; WANG et al., 2015), com $1 \geq N \leq \infty$.

Devido à complexidade computacional das técnicas determinísticas, houve uma concentração de trabalhos nas técnicas de AMS, que podem ser baseadas em diversas heurísticas (EDGAR; BATZOGLOU, 2006). Destas, destaca-se atualmente o Alinhamento Progressivo (FENG; DOOLITTLE, 1987), que é o mais implementado nas diversas ferramentas bem conhecidas, como as da família Clustal (THOMPSON; HIG-

GINS; GIBSON, 1994; SIEVERS; HIGGINS, 2014; SIEVERS; HIGGINS, 2018), a MUSCLE (EDGAR, 2004a), a MAFFT (KATOH; ROZEWICKI; YAMADA, 2017), a MULTAL (TAYLOR, 1988), entre outras.

A característica que torna o Alinhamento Progressivo interessante deve-se ao fato de que as sequências mais semelhantes, ou menos distantes, são alinhadas primeiro por meio de uma árvore binária, que neste contexto recebe o nome de árvore guia. A árvore guia é construída com auxílio de uma matriz de distâncias, na qual são armazenadas as distâncias evolutivas computadas com o auxílio de uma função de medida, para todos os possíveis pares de sequências de um dado conjunto de sequências fornecido como entrada para o algoritmo. Os pares de sequências menos distantes, ou mais semelhantes, são agrupados primeiro e assume-se que, no Alinhamento Progressivo, o melhor resultado é obtido em cada nó ao alinhar os dois perfis com menor número de diferenças, mesmo que não sejam vizinhos evolutivos (BOYCE; SIEVERS; HIGGINS, 2015; EDGAR, 2004b).

O Alinhamento Progressivo tem como vantagem a velocidade, a simplicidade e a sensibilidade. Porém, erros ocorridos no início do processo não são corrigidos nas fases posteriores, devido à característica gulosa da estratégia e podem conduzir a distorções no resultado final (NOTREDAME; HOLM; HIGGINS, 1998). Isto leva à necessidade de esforços para melhorar a acurácia da árvore guia. O desenvolvimento de estratégias capazes de conciliar desempenho computacional e acurácia dos alinhamentos, dando um significado biológico relevante é um dos desafios na área da Bioinformática (MORRISON; MORGAN; KELCHNER, 2015). Uma vez que o Alinhamento Progressivo é sensível aos problemas que ocorrem na fase de cálculo das distâncias, as técnicas de rearranjo da árvore guia destacaram-se como um ponto fundamental, pois podem amenizar essa distorção (HSIEH et al., 2015).

Neste contexto, observam-se vários estudos apontando para a importância da árvore guia e para a necessidade de melhorar sua acurácia, de modo a garantir um ali-

nhamento final com maior significância biológica. Boyce (BOYCE; SIEVERS; HIGGINS, 2015) demonstrou que uma simples mudança na ordem dos conjuntos de entrada, pode produzir resultados diferentes para esse mesmo conjunto. Penn (PENN et al., 2010) mostrou que as incertezas na árvore guia são fontes importantes de incertezas no alinhamento. Capella-Gutierrez e Gabaldon (CAPELLA-GUTIÉRREZ; GABALDÓN, 2013) mostraram que a maioria das lacunas (**gaps**) são inseridas em padrões que seguem a árvore guia. Zhan (ZHAN et al., 2015) usou um método adaptativo de construção da árvore para demonstrar a melhora da acurácia de diversas ferramentas de AMS com o uso de melhores árvores guia. Algumas ferramentas também têm obtido melhores resultados por meio do uso de técnicas para construir melhores árvores guia, como GLProbs (YE et al., 2015) e a PnpProbs (YE; LAM; TING, 2016). Hsieh (HSIEH et al., 2015) usou uma adaptação da técnica de bissecção e reconexão da árvore *(TBR - Tree Bisection and Reconnection)*, a fim de melhorar a reconstrução de árvores guia, combinada com o princípio da evolução mínima *(ME - Minimum Evolution Principle)* (RZHETSKY; NEI, 1993) para filtrar posições reconectadas desnecessárias e reduzir o tempo de busca, demonstrando que o método pode ajudar outros algoritmos na construção de árvores mais precisas, com tempo de processamento aceitável.

1.3 Motivação e Escopo

A partir dos estudos citados na seção 1.2 demonstra-se a importância da árvore guia, também conhecida como árvore filogenética, que é produzida pelo AMS durante a execução da segunda fase do processo de alinhamento. Melhorar a sua acurácia, ou seja, aplicar técnicas que promovam melhorias a fim de gerar uma árvore de tamanho mínimo e, que desta forma, reflita com maior proximidade a linha da evolução das espécies, ainda é um problema na área da Bioinformática. Uma árvore guia mais acurada é uma árvore em que as sequências com a menor distância evolutiva estão agrupadas juntas, ou mais próximas. Dessa forma, pode-se gerar arestas com pesos

mínimos, a fim de representar melhor a evolução das espécies e, consequentemente, são capazes de produzir um resultado do AMS com maior significância biológica.

Técnicas que promovem mudanças na topologia da árvore guia tem por objetivo produzir uma árvore de tamanho mínimo, para a qual a soma dos tamanhos de todas as suas arestas, aproxima-se ao máximo do tamanho da árvore evolutiva verdadeira, que é uma árvore que representa a verdadeira evolução das espécies. Neste contexto, a acurácia de uma árvore guia é medida pelo resultado da diferença de seu tamanho quando comparado com o tamanho da árvore verdadeira, onde quanto menor a diferença, melhor é a árvore produzida.

Produzir resultados do AMS com maior significância biológica, com um resultado capaz de representar com maior proximidade a evolução das espécies e demonstrar suas similaridades, o mais próximo da verdadeira evolução, é um fator importante para os pesquisadores das áreas de saúde, da indústria farmacêutica e das ciências biológicas. Na área de saúde e na indústria farmacêutica, tais resultados são importantes para determinar quais são as terapias e medicamentos mais adequados para determinada enfermidade ou no combate de vírus e bactérias nocivos a saúde.

Sendo assim, constitui-se como motivação deste trabalho desenvolver um método capaz interagir com as ferramentas de AMS e promover melhorias na árvore guia, utilizando técnicas de rearranjo de árvore, convergindo para resultados finais do AMS com maior significância biológica e contribuir com os trabalhos de biólogos, pesquisadores e profissionais de saúde.

1.4 Objetivos

O presente trabalho tem por objetivo o desenvolvimento de um método de rearranjo que promova melhorias na acurácia da árvore guia e produza uma árvore guia com tamanhos mínimos de arestas, mais próxima da árvore evolutiva verdadeira, gerando um resultado com maior qualidade nos alinhamentos produzidos, sem degradar o de-

sempenho computacional em termos de tempo de execução e, assim, contribuir para a evolução da significância biológica dos resultados finais produzidos pelo AMS.

O método atua nas ferramentas de AMS entre as fases de construção da árvore guia e de construção do AMS, mas não se limita a uma única ferramenta. Desta forma, foi desenvolvido em um módulo separado que recebe como entrada uma árvore guia e sua matriz de distâncias, executa trocas em suas ramificações internas para melhorar a sua acurácia e retorna a árvore guia modificada para a ferramenta de AMS. Para o desenvolvimento do presente trabalho, foi escolhida uma única ferramenta que serviu de base para medição dos resultados, que, neste caso, foi a ferramenta MUSCLE (EDGAR et al., 2005).

Sendo assim, trata-se de uma estratégia de rearranjo de árvores com o propósito de produzir árvores guias vizinhas da árvore guia inicial, produzida pela ferramenta de alinhamento hospedeira. Desta forma, é possível dota-la da capacidade de selecionar uma árvore melhor e de tamanho reduzido, mais próximo da filogenia verdadeira. Com isso, majora-se a capacidade de produzir melhores resultados no alinhamento das sequências biológicas, com o uso do Alinhamento Progressivo.

1.5 Organização do Trabalho

O presente trabalho está dividido em cinco capítulos, incluindo o atual. No Capítulo 2, realiza-se uma revisão bibliográfica referente aos assuntos e conceitos pertinentes a área da Biologia e da Bioinformática, bem como apresenta-se algumas das técnicas de AMS e de rearranjo de árvores pertinentes ao escopo do trabalho da dissertação desenvolvida. No capítulo 3, apresenta-se a proposta do método para rearranjo da árvore guia e todo o trabalho executado para o seu desenvolvimento. No capítulo 4, os resultados obtidos como o uso do método são medidos e comprados com os resultados obtidos pela ferramenta sem o uso do método proposto. Por fim, no Capítulo 5, apresentam-se as conclusões do presente trabalho e os trabalhos futuros.

Capítulo 5

Conclusões

5.1 Considerações Finais

No presente trabalho, para o bom entendimento do contexto no qual a Bioinformática está inserida, apresentamos vários conceitos fundamentais e uma breve descrição das principais ferramentas e heurísticas para o Alinhamento Múltiplo de Sequências, com maior atenção para a heurística do Alinhamento Progressivo. Também apresentamos os conceitos básicos para o entendimento do contexto do problema da reconstrução filogenética e rearranjo de árvores. Destacou-se várias técnicas aplicadas à solução do referido problema, com objetivo de obter-se melhores resultados no Alinhamento Múltiplo de Sequência. Uma atenção especial foi dada à técnica TBR, a fim de apresentar um resultado final do alinhamento com maior significância biológica.

Sendo assim, buscou-se apresentar os bons resultados obtidos pelo MUSCLE, uma das principais e mais utilizadas ferramentas na área de Bioinformática e o funcionamento da TBR, bem como os bons resultados obtidos da sua utilização. Além disso, destacou-se os pontos em que se aplicou a TBR na MUSCLE, a fim de melhorar os resultados dos alinhamentos de sequencias produzidos.

A escolha da TBR deu-se pelos excelentes resultados apresentados em sua aplicação por Hsieh et al. (HSIEH et al., 2015). A escolha do MUSCLE deu-se pela ampla

utilização da ferramenta em Bioinformática, por sua abordagem computacional bem estruturada e sua capacidade de produzir bons resultados com custo computacional baixo (EDGAR; BATZOGLOU, 2006). Além do que, os códigos-fonte da MUSCLE são abertos e encontram-se muito bem fundamentados no paradigma de programação Orientado a Objetos, tornando o desenvolvimento da proposta mais robusto.

A aplicação do TreeRMEP para reconstrução filogenética produzida pela função padrão do MUSCLE, logo após a geração da primeira árvore guia, sem refinamentos, foi capaz de obter uma árvore guia melhor, com distâncias menores entre suas ramificações, de acordo com o Principio da Evolução Miníma. Isto resultou em uma árvore filogenética mais próxima da árvore verdadeira, resultando em contribuição para produzir alinhamentos de sequências com maior significância biológica.

Os resultados obtidos comprovaram a contribuição do método, com pequeno aumento no tempo computacional, o que era esperado, mas ficou dentro de níveis que não inviabilizam o método, o que é muito importante, uma vez que existe a necessidade de evolução na capacidade da MUSCLE de operar com grandes conjuntos de dados, especialmente os produzidos por NGS, o que é um fator de limitação da ferramenta.

5.2 Trabalhos Futuros

Apesar do tempo de execução ter ficado em limites aceitáveis, existe a possibilidade do desenvolvimento de uma versão do método para ambientes paralelizados, como *grids* ou *clusters*. Atualmente muitas aplicações tem usufruído do poder de processamento em GPU *(Graphics Processing Unit)* e uma versão do TreeRMEP paralelizado será capaz de compensar o acréscimo no tempo de execução.

Outro ponto é que para algumas sequências o método não foi capaz de encontrar uma árvore melhorada, o que abre a possibilidade para se investigar outras funções de medidas capazes de avaliar com maior precisão as topologias geradas e, como isso, aumentar a qualidade dos resultados obtidos pelas ferramentas de AMS.

REFERÊNCIAS

ADAMS, R. L. P. et al. *The biochemistry of the nucleic acids*. [S.l.]: Chapman and Hall, 1992.

ALTSCHUL, S. F.; CARROLL, R. J.; LIPMAN, D. J. Weights for data related by a tree. *Journal of molecular biology*, Elsevier, v. 207, n. 4, p. 647–653, 1989.

AMARAN, S. et al. Simulation optimization: a review of algorithms and applications. *Annals of Operations Research*, Springer, v. 240, n. 1, p. 351–380, 2016.

BAJORATH, J.; STENKAMP, R.; ARUFFO, A. Knowledge-based model building of proteins: concepts and examples. *Protein science: a publication of the Protein Society*, Blackwell Publishing, v. 2, n. 11, p. 1798, 1993.

BASTKOWSKI, S. et al. The minimum evolution problem is hard: a link between tree inference and graph clustering problems. *Bioinformatics*, Oxford University Press, v. 32, n. 4, p. 518–522, 2015.

BEHJATI, S.; TARPEY, P. S. What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, Royal College of Paediatrics and Child Health, v. 98, n. 6, p. 236–238, 2013.

BLACKSHIELDS, G. et al. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms for Molecular Biology*, BioMed Central, v. 5, n. 1, p. 21, 2010.

BORDEWICH, M. et al. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, IEEE Computer Society Press, v. 6, n. 1, p. 110–117, 2009.

BORDEWICH, M.; SEMPLE, C. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of combinatorics*, Springer, v. 8, n. 4, p. 409–423, 2005.

BOYCE, K.; SIEVERS, F.; HIGGINS, D. G. Instability in progressive multiple sequence alignment algorithms. *Algorithms for molecular biology*, BioMed Central, v. 10, n. 1, p. 26, 2015.

BRAZMA, A. et al. Approaches to the automatic discovery of patterns in biosequences. *Journal of computational biology*, v. 5, n. 2, p. 279–305, 1998.

BRYANT, D. The splits in the neighborhood of a tree. *Annals of Combinatorics*, Springer, v. 8, n. 1, p. 1–11, 2004.

BUSIA, A. et al. A deep learning approach to pattern recognition for short dna sequences. *bioRxiv*, Cold Spring Harbor Laboratory, p. 353474, 2018.

CAPELLA-GUTIÉRREZ, S.; GABALDÓN, T. Measuring guide-tree dependency of inferred gaps in progressive aligners. *Bioinformatics*, Oxford University Press, v. 29, n. 8, p. 1011–1017, 2013.

CHAPIN, E. A.; CHAPIN, C. C. Darwin and the galapagos islands. *Bull. Pan Am. Union*, HeinOnline, v. 69, p. 655, 1935.

CHATZOU, M. et al. Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, Oxford University Press, v. 17, n. 6, p. 1009–1023, 2015.

COHEN, J. Bioinformatics—an introduction for computer scientists. *ACM Computing Surveys (CSUR)*, ACM, v. 36, n. 2, p. 122–158, 2004.

DAHM, R. Discovering dna: Friedrich miescher and the early years of nucleic acid research. *Human genetics*, Springer, v. 122, n. 6, p. 565–581, 2008.

DAY, W. H. Properties of the nearest neighbor interchange metric for trees of small size. *Journal of Theoretical Biology*, Elsevier, v. 101, n. 2, p. 275–288, 1983.

DENIS, F.; GASCUEL, O. On the consistency of the minimum evolution principle of phylogenetic inference. *Discrete Applied Mathematics*, Elsevier, v. 127, n. 1, p. 63–77, 2003.

DESPER, R.; GASCUEL, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of computational biology*, Mary Ann Liebert, Inc., v. 9, n. 5, p. 687–705, 2002.

DORIGO, M.; BLUM, C. Ant colony optimization theory: A survey. *Theoretical computer science*, Elsevier, v. 344, n. 2, p. 243–278, 2005.

EDGAR, R. C. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, BioMed Central, v. 5, n. 1, p. 113, 2004.

EDGAR, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, Oxford University Press, v. 32, n. 5, p. 1792–1797, 2004.

EDGAR, R. C.; BATZOGLOU, S. Multiple sequence alignment. *Current opinion in structural biology*, Elsevier, v. 16, n. 3, p. 368–373, 2006.

EDGAR, R. C. et al. *MUSCLE user guide*. [S.l.], 2005.

EIDHAMMER, I.; JONASSEN, I.; TAYLOR, W. R. Structure comparison and structure patterns. *Journal of Computational Biology*, Mary Ann Liebert, Inc., v. 7, n. 5, p. 685–716, 2000.

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

FILIPSKI, A. et al. Phylogenetic placement of metagenomic reads using the minimum evolution principle. *BMC genomics*, BioMed Central, v. 16, n. 1, p. S13, 2015.

GAO, X.; ZHANG, J.; WEI, Z. Deep learning for sequence pattern recognition. In: IEEE. *Networking, Sensing and Control (ICNSC), 2018 IEEE 15th International Conference on*. [S.l.], 2018. p. 1–6.

GEST, H. The discovery of microorganisms by robert hooke and antoni van leeuwenhoek, fellows of the royal society. *Notes and records of the Royal Society of London*, The Royal Society, v. 58, n. 2, p. 187–201, 2004.

GIRIBET, G. Efficient tree searches with available algorithms. *Evolutionary bioinformatics online*, SAGE Publications, v. 3, p. 341, 2007.

GLENN, T. C. Field guide to next-generation dna sequencers. *Molecular ecology resources*, Wiley Online Library, v. 11, n. 5, p. 759–769, 2011.

GLOVER, F.; LAGUNA, M. *Tabu search*. [S.l.]: Springer, 1999.

GOLDBERG, D. E.; HOLLAND, J. H. Genetic algorithms and machine learning. *Machine learning*, Springer, v. 3, n. 2, p. 95–99, 1988.

GONDRO, C.; KINGHORN, B. A simple genetic algorithm for multiple sequence alignment. *Genetics and Molecular Research*, v. 6, n. 4, p. 964–982, 2007.

GOTOH, O. An improved algorithm for matching biological sequences. *Journal of molecular biology*, Elsevier, v. 162, n. 3, p. 705–708, 1982.

GOULD, S. J. *Ontogeny and phylogeny*. [S.l.]: Harvard University Press, 1977.

GUERRA, M. K. M.; BUCKLER, E. S. k-mer grammar uncovers maize regulatory architecture. *bioRxiv*, Cold Spring Harbor Laboratory, p. 222927, 2017.

GUINDON, S.; GASCUEL, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, Society of Systematic Zoology, v. 52, n. 5, p. 696–704, 2003.

GUSFIELD, D. *Algorithms on strings, trees and sequences: computer science and computational biology*. [S.l.]: Cambridge university press, 1997.

HIGGINS, D. G.; SHARP, P. M. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, Elsevier, v. 73, n. 1, p. 237–244, 1988.

HIROSAWA, M. et al. Comprehensive study on iterative algorithms of multiple sequence alignment. *Bioinformatics*, Oxford University Press, v. 11, n. 1, p. 13–18, 1995.

HSIEH, S.-Y. et al. An enhanced algorithm for reconstructing a phylogenetic tree based on the tree rearrangement and maximum likelihood method. In: SPRINGER. *International Conference on Intelligent Computing*. [S.l.], 2015. p. 530–541.

KATOH, K. et al. Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*, Oxford University Press, v. 33, n. 2, p. 511–518, 2005.

KATOH, K. et al. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, Oxford Univ Press, v. 30, n. 14, p. 3059–3066, 2002.

KATOH, K.; ROZEWICKI, J.; YAMADA, K. D. Mafft online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*, 2017.

KATOH, K.; STANDLEY, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, SMBE, v. 30, n. 4, p. 772–780, 2013.

KAUR, H.; CHAND, L. Biological sequence alignment using varied optimization algorithms. In: IEEE. *Inventive Computation Technologies (ICICT), International Conference on*. [S.l.], 2016. v. 1, p. 1–5.

KHURI, S. A bioinformatics track in computer science. In: ACM. *ACM SIGCSE Bulletin*. [S.l.], 2008. v. 40, n. 1, p. 508–512.

KIDD, K. K.; SGARAMELLA-ZONTA, L. A. Phylogenetic analysis: concepts and methods. *American journal of human genetics*, Elsevier, v. 23, n. 3, p. 235, 1971.

KIRKPATRICK, S. et al. Optimization by simulated annealing. *science*, World Scientific, v. 220, n. 4598, p. 671–680, 1983.

KOCEV, D. et al. Tree ensembles for predicting structured outputs. *Pattern Recognition*, Elsevier, v. 46, n. 3, p. 817–833, 2013.

LASSMANN, T.; SONNHAMMER, E. L. Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, BioMed Central, v. 6, n. 1, p. 298, 2005.

LEMOS, M.; ARAGAO, M. V. S. P.; CASANOVA, M. A. *Padrões em Biossequências*. [S.l.]: PUC, 2003.

LEMOS, M.; BASÍLIO, A.; CASANOVA, M. A. *Um estudo dos algoritmos de montagen de fragmentos de DNA*. [S.l.]: PUC, 2003.

LEMOS, M.; CASANOVA, M. A. *Algoritmos para análise de sequências*. [S.l.]: PUC, 2000.

LI, M.; TROMP, J.; ZHANG, L. On the nearest neighbour interchange distance between evolutionary trees. *Journal of Theoretical Biology*, Elsevier, v. 182, n. 4, p. 463–467, 1996.

LIEW, A. W.-C.; YAN, H.; YANG, M. Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recognition*, Elsevier, v. 38, n. 11, p. 2055–2073, 2005.

LIU, L. et al. Comparison of next-generation sequencing systems. *BioMed Research International*, Hindawi Publishing Corporation, v. 2012, 2012.

LIU, Y.; SCHMIDT, B.; MASKELL, D. L. Msaprobs: multiple sequence alignment based on pair hidden markov models and partition function posterior probabilities. *Bioinformatics*, Oxford University Press, v. 26, n. 16, p. 1958–1964, 2010.

LÓPEZ-IBÁÑEZ, M.; STÜTZLE, T.; DORIGO, M. Ant colony optimization: A component-wise overview. *Handbook of Heuristics*, Springer, p. 1–37, 2016.

MOORE, R. E.; YOUNG, M. K.; LEE, T. D. Qscore: an algorithm for evaluating sequest database search results. *Journal of the American Society for Mass Spectrometry*, Springer, v. 13, n. 4, p. 378–386, 2002.

MORGENSTERN, B. Multiple sequence alignment with dialign. In: *Multiple Sequence Alignment Methods*. [S.l.]: Springer, 2014. p. 191–202.

MORGENSTERN, B. et al. Dialign: finding local similarities by multiple sequence alignment. *Bioinformatics*, Oxford Univ Press, v. 14, n. 3, p. 290–294, 1998.

MORRISON, D.; MORGAN, M.; KELCHNER, S. Molecular homology and multiple sequence alignment: an analysis of concepts and practice. *Australian Systematic Botany*, CSIRO, 2015.

MOSS, J.; JOHNSON, C. G. An ant colony algorithm for multiple sequence alignment in bioinformatics. In: SPRINGER. *Artificial Neural Nets and Genetic Algorithms*. [S.l.], 2003. p. 182–186.

NAIDU, V.; NARAYANAN, A. Needleman-wunsch and smith-waterman algorithms for identifying viral polymorphic malware variants. In: IEEE. *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. [S.l.], 2016. p. 326–333.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, Elsevier, v. 48, n. 3, p. 443–453, 1970.

NELESEN, S. M. et al. The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analysis. In: *Pacific Symposium on Biocomputing*. [S.l.: s.n.], 2008. v. 13, n. 2008, p. 25–36.

NOTREDAME, C.; HIGGINS, D. G. Saga: sequence alignment by genetic algorithm. *Nucleic acids research*, Oxford Univ Press, v. 24, n. 8, p. 1515–1524, 1996.

NOTREDAME, C.; HIGGINS, D. G.; HERINGA, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, Elsevier, v. 302, n. 1, p. 205–217, 2000.

NOTREDAME, C.; HOLM, L.; HIGGINS, D. G. Coffee: an objective function for multiple sequence alignments. *Bioinformatics*, Oxford Univ Press, v. 14, n. 5, p. 407–422, 1998.

OGDEN, T. H.; ROSENBERG, M. S. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic biology*, Society of Systematic Zoology, v. 55, n. 2, p. 314–328, 2006.

PAL, S. K.; WANG, P. P. *Genetic algorithms for pattern recognition*. [S.l.]: CRC press, 2017.

PAPADIMITRIOU, C. H.; STEIGLITZ, K. *Combinatorial optimization: algorithms and complexity*. [S.l.]: Courier Corporation, 1998.

PENN, O. et al. An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular Biology and Evolution*, Oxford University Press, v. 27, n. 8, p. 1759–1767, 2010.

PROSDOCIMI, F. et al. Bioinformática: manual do usuário. *Biotecnologia Ciência & Desenvolvimento*, v. 29, p. 12–25, 2002.

REEVES, C. R. Genetic algorithms for the operations researcher. *INFORMS journal on computing*, INFORMS, v. 9, n. 3, p. 231–250, 1997.

RIAZ, T.; WANG, Y.; LI, K.-B. Multiple sequence alignment using tabu search. In: AUSTRALIAN COMPUTER SOCIETY, INC. *Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29*. [S.l.], 2004. p. 223–232.

RIDDER, D. de; RIDDER, J. de; REINDERS, M. J. Pattern recognition in bioinformatics. *Briefings in bioinformatics*, Oxford University Press, v. 14, n. 5, p. 633–647, 2013.

RUBIN, G. M. et al. Comparative genomics of the eukaryotes. *Science*, American Association for the Advancement of Science, v. 287, n. 5461, p. 2204–2215, 2000.

RZHETSKY, A.; NEI, M. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular biology and evolution*, v. 10, n. 5, p. 1073–1095, 1993.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, v. 4, n. 4, p. 406–425, 1987.

SCHWEFEL, H.-P.; RUDOLPH, G. *Contemporary evolution strategies*. [S.l.]: Springer, 1995.

SIEVERS, F.; HIGGINS, D. G. Clustal omega, accurate alignment of very large numbers of sequences. In: *Multiple sequence alignment methods*. [S.l.]: Springer, 2014. p. 105–116.

SIEVERS, F.; HIGGINS, D. G. Clustal omega for making accurate alignments of many protein sequences. *Protein Science*, Wiley Online Library, v. 27, n. 1, p. 135–145, 2018.

SIEVERS, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, EMBO Press, v. 7, n. 1, p. 539, 2011.

SLATKIN, M.; MADDISON, W. P. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, Genetics Soc America, v. 123, n. 3, p. 603–613, 1989.

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

SOKAL, R. R. A statistical method for evaluating systematic relationship. *University of Kansas science bulletin*, v. 28, p. 1409–1438, 1958.

STALLINGS, W. *Computer organization and architecture: designing for performance*. [S.l.]: Pearson Education India, 2003.

STAMATAKIS, A.; HOOVER, P.; ROUGEMONT, J. A rapid bootstrap algorithm for the raxml web servers. *Systematic biology*, Taylor & Francis, v. 57, n. 5, p. 758–771, 2008.

TAYLOR, W. R. A flexible method to align large numbers of biological sequences. *Journal of Molecular Evolution*, Springer, v. 28, n. 1-2, p. 161–169, 1988.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, Oxford Univ Press, v. 22, n. 22, p. 4673–4680, 1994.

THOMPSON, J. D. et al. Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 61, n. 1, p. 127–136, 2005.

TURING, A. M. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London mathematical society*, Wiley Online Library, v. 2, n. 1, p. 230–265, 1937.

WALLACE, I. M.; BLACKSHIELDS, G.; HIGGINS, D. G. Multiple sequence alignments. *Current opinion in structural biology*, Elsevier, v. 15, n. 3, p. 261–266, 2005.

WANG, L.; JIANG, T. On the complexity of multiple sequence alignment. *Journal of computational biology*, v. 1, n. 4, p. 337–348, 1994.

WANG, X.-D. et al. A survey of multiple sequence alignment techniques. In: SPRINGER. *International Conference on Intelligent Computing*. [S.l.], 2015. p. 529–538.

WATERMAN, M.; SMITH, T.; BEYER, W. Some biological sequence metrics. *Advances in Mathematics*, v. 20, n. 3, p. 367 – 387, 1976. ISSN 0001-8708. Disponível em: <http://www.sciencedirect.com/science/article/pii/0001870876902024>.

WATSON, J. D.; CRICK, F. H. et al. Molecular structure of nucleic acids. *Nature*, v. 171, n. 4356, p. 737–738, 1953.

WHITMAN, W. B.; COLEMAN, D. C.; WIEBE, W. J. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 95, n. 12, p. 6578–6583, 1998.

WU, Y. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, Oxford University Press, v. 25, n. 2, p. 190–196, 2008.

XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science*, Springer, v. 2, n. 2, p. 165–193, 2015.

YE, Y. et al. Glprobs: Aligning multiple sequences adaptively. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, IEEE Computer Society Press, v. 12, n. 1, p. 67–78, 2015.

YE, Y.; LAM, T.-W.; TING, H.-F. Pnpprobs: a better multiple sequence alignment tool by better handling of guide trees. *BMC bioinformatics*, BioMed Central, v. 17, n. 8, p. 285, 2016.

ZAFALON, G. et al. A parallel approach of coffee objective function to multiple sequence alignment. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2015. v. 633, n. 1, p. 012084.

ZAFALON, G. F. D. *Aplicação de estratégias híbridas em algoritmos de alinhamento múltiplo de sequências para ambientes de computação paralela e distribuída.* Tese (Doutorado) — Universidade de São Paulo, 2014.

ZHAN, Q. et al. Improving multiple sequence alignment by using better guide trees. *BMC bioinformatics*, BioMed Central, v. 16, n. 5, p. S4, 2015.

ZHU, X. et al. Parallel implementation of mafft on cuda-enabled graphics hardware. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, IEEE, v. 12, n. 1, p. 205–218, 2015.