Estimativa de Tempo de Alinhamentos Biológicos na Nuvem

Gabriel C. Mills[1], **Sara M. Cavalcante**[1], **Cristina Boeres**[1], **Vinod E.F. Rebello**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brazil

{gmills,saracavalcante}@id.uff.br, {boeres,vinod}@ic.uff.br

Abstract. *Cloud computing's efficiency and lower costs are dependent on effective resource management. One key aspect is predicting application execution times in order to define good schedules. This paper analyzes the runtime predictability of a cloud service for DNA sequence alignments designed to efficiently utilize cloud resources to allow scientists to evaluate the thousands of alignments that may form a single experiment. However, the quality of the resource allocation is based on the assumption that all of the alignments in the experiment take the same amount of time. This work aims to identify factors and their degree of impact on the execution time of each alignment.*

Resumo. *A eficiência e o baixo custo da Computação em Nuvem dependem do gerenciamento de recursos computacionais. Um aspecto vital dessa tarefa é a estimativa do tempo de execução a fim de definir bons escalonamentos. Este artigo analisa a previsibilidade desse tempo de um serviço em nuvem para o alinhamento de sequências de DNA projetado com o fim de utilizar recursos eficientemente e viabilizar a análise pelos cientistas dos milhares de alinhamentos que podem compor um único experimento. Entretanto, a qualidade da alocação de recursos é baseada na hipótese de que todos os alinhamentos do experimento demandam o mesmo tempo. Este trabalho visa à análise de fatores e os respectivos graus com que impactam o tempo de execução de cada alinhamento.*

1. Introdução

A realização de alinhamentos de sequências de DNA ou RNA para estudar, por exemplo, novas variantes de um vírus é essencial para entender a infecção que este pode causar, a facilidade com que se espalha, a gravidade dos sintomas e a eficácia das vacinas. A recente pandemia apenas destacou a importância desse tipo de análise, como exemplificado pelos mais de 18 milhões de sequências genéticas do vírus SARS-CoV-2 disponíveis para estudo por cientistas, ao redor do mundo, em bancos de dados públicos, como do *GenBank* do *National Center for Biotechnology Information* (NCBI) [Sayers 2022].

O alinhamento de sequências é uma das áreas chave da Bioinformática responsável pela identificação da similaridade de sequências genéticas. Embora existam serviços em nuvem de alinhamento (como o *Basic Local Alignment Search Tool* do NCBI, hospedado no Amazon EC2), a maioria tem limites restritivos em comprimentos das sequências e pouco se sabe sobre a eficiência de suas implementações. O trabalho de [Sodré et al. 2022] especificou um serviço de alinhamento em nuvem, comprovadamente eficiente em termos de qualidade da solução, tempo total de execução e custos financeiros. Ainda, a ferramenta *Multi-Platform Architecture for Sequence Aligners*

(MASA) foi adotada devido a sua capacidade de realizar alinhamentos de sequências com mais de 200 milhões de nucleotídeos, em uma variedade de plataformas de hardware/software [De O. Sandes et al. 2016]. A versão MASA-OpenMP foi projetada para aproveitar os vários núcleos em um servidor e, portanto, seria capaz de aproveitar a gama de arquiteturas oferecidas por provedores de nuvem. MASA alinha duas sequências de DNA ou RNA usando uma variação [Myers and Miller 1988] do algoritmo clássico de Smith-Waterman para encontrar o alinhamento *ótimo*. Com uma complexidade quadrática de tempo e linear do espaço de memória, MASA emprega um método de *pruning* para tentar reduzir custos computacionais de alinhamentos [De O. Sandes et al. 2016].

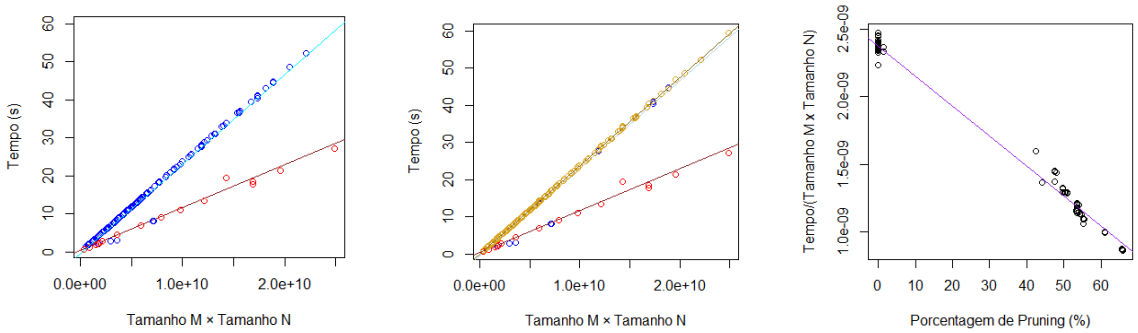
A adoção da computação em nuvem cresce rapidamente, em especial, o modelo *Infrastructure as a Service*, no qual provedores oferecem seus recursos na forma de *instâncias virtuais*, com diferentes capacidades pré-configuradas a determinadas taxas de custo por hora. Foi proposta uma técnica para escalonar vários alinhamentos em uma única instância [Sodré et al. 2022], sendo tal escalonamento praticamente ótimo e que efetivamente dobra o rendimento dos fluxos de trabalho de alinhamento de sequências feito por MASA-OpenMP. Contudo, o modelo pressupõe que os alinhamentos do mesmo experimento levem tempos de execução iguais. Para entender a robustez da solução, é necessário avaliar o quanto os tempos podem variar. Este artigo investiga experimentalmente tal variabilidade executando MASA na nuvem.

2. Descrição do experimento

Um conjunto de sequências variadas de DNA foi coletado, em formato FASTA, do banco de dados *GenBank* de sequências genéticas do NCBI [Sayers 2022]. Dois *bash scripts* foram utilizados, diferenciados apenas pelo argumento `'--no-block-pruning'` no comando executado, com o fim de medir o tempo de execução e, quando habilitada, a taxa de *pruning* alcançada ao alinhar os pares de sequências, inclusive um par da mesma sequência. Os tempos apresentados são médias de três execuções. A análise foi realizada no ambiente de nuvem EC2 (*Amazon Elastic Compute Cloud*) na região us-east-1, utilizando a instância otimizada para computação C7g.xlarge, com processadores AWS Graviton 3, 4 CPUs, 8 GiB de memória RAM e um volume SSD do tipo GP3. O MASA-OpenMP foi executado em thread única no sistema Linux Ubuntu Jammy 22.04.

3. Análise dos tempos de execução

A análise investigou se algoritmos de alinhamento requerem um tempo de execução proporcional ao produto dos comprimentos das duas sequências e se MASA com *pruning* alcança tempos menores. As Figuras 1(a) e 1(b) apresentam os tempos de execução de MASA em relação ao produto $m \times n$, em que m e n são os números de nucleotídeos nas duas respectivas sequências sendo alinhadas. A Figura 1(a) divide os tempos de MASA com *pruning* em dois grupos: os pontos sobre a reta vermelha (com coeficiente angular 1.124e-09) de pares de sequências iguais e os pontos sobre a reta azul (com coeficiente angular 2.356e-09) de pares de sequências diferentes. Observa-se que o tempo de execução é proporcional ao produto dos comprimentos para os dois grupos. Vê-se também que há alinhamentos com *pruning* de sequências distintas, porém com um alto grau de similaridade entre si, cujo tempo está sobre a reta vermelha, como a aqueles entre variantes de SARS-CoV-2. No entanto, existem casos em que *pruning* não reduz o tempo. Isso é ratificado pela Figura 1(b) por meio da sobreposição dos tempos, para todos os mesmos



(a) Tempos para alinhar sequências distintas (azul) e iguais (vermelhas) usando *pruning*

(b) Tempos sem *pruning* para todas sequências (amarelo), sobrepostos aos tempos da Figura 1(a)

(c) Tempo normalizado pelo produto do tamanho das sequências contra a taxa de *pruning*

Figura 1. Tempos (s) do MASA considerando diversos pares de sequências

alinhamentos da Figura 1(a), alcançados por MASA sem usar *pruning*. A reta amarela tem um coeficiente angular de 2.374e-09, quase idêntico ao da reta azul. O fato de que nenhum ponto amarelo aparece próximo à reta vermelha indica que *pruning* pode alcançar reduções de até 75% do tempo quando existe um grau de similaridade significativo entre as sequências, cenário mais comum em avaliações científicas. A Figura 1(c) tenta quantificar a relação do *pruning* na diminuição do tempo de execução. Em geral, quanto maior a porcentagem de *pruning*, maior a redução de tempo. Dadas a equação da reta roxa e a taxa de *pruning*, seria possível obter uma boa estimativa do tempo de execução.

4. Conclusão e análises futuras

O alinhamento de duas sequências é uma etapa importante na solução de vários problemas na Bioinformática. Este artigo apresentou uma análise inicial da ferramenta MASA sendo utilizada em um serviço de nuvem. O baixo custo do serviço necessita que os alinhamentos de um mesmo experimento levem tempos similares. Este trabalho almeja estimar tais tempos e a análise confirmou que estes dependem do produto dos comprimentos das duas sequências sendo alinhadas. A taxa de *pruning* alcançada melhora o tempo de execução, mas varia em relação às caraterísticas das sequências. Trabalhos futuros pretendem identificar a relação entre a similaridade das sequências e a taxa de *pruning* alcançada.

Referências

De O. Sandes, E. F., Miranda, G., Martorell, X., Ayguade, E., Teodoro, G., and De Melo, A. C. M. A. (2016). MASA: A multiplatform architecture for sequence aligners with block pruning. *ACM Transactions on Parallel Computing*, 2(4).

Myers, E. W. and Miller, W. (1988). Optimal alignments in linear space. *Bioinformatics*, 4(1):11–17.

Sayers, E. W. *et al.*. (2022). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 50(D1):D20–D26.

Sodré, D., Boeres, C., and Rebello, V. (2022). Making the most of what you pay for by delaying tasks to improve overall cloud instance performance. In *Anais Estendidos do XXIII Simpósio em Sistemas Computacionais de Alto Desempenho*, pages 9–16. SBC.