PA-Star-Web: *web server* para obtenção do alinhamento múltiplo ótimo de sequências biológicas

Enéias Paulo de Oliveira[1], **Daniel Sundfeld**[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Brasília (IFB)
Campus Brasília, Asa Norte – 70.830-450 – Brasília – DF – Brasil

[2]Faculdade UnB Gama (FGA) – Universidade de Brasília (UnB)
Campus Gama, Setor Leste – 72.444-240 – Gama – DF – Brasil

eneiaspaulo1984@gmail.com, daniel.sundfeld@unb.br

Abstract. *Sequence comparison is one of the most important operations in Bioinformatics. When comparing multiple sequences, the problem becomes complex and requires a high amount of memory and time. PA-Star is a tool that achieves the optimal multiple alignment of multiple sequences. However, this tool produces results exclusively through the command line interface. In this work, we propose a web server to provide access to results with a web browser. The web server is deployed on multiple virtual machines on a cloud computing provider and they can be scaled up to fit the workload.*

Resumo. *A comparação de sequências é uma das operações mais importantes da Bioinformática. Ao se comparar múltiplas sequências, o problema torna-se complexo e exige uma alta quantidade de memória e tempo computacional. O PA-Star é uma ferramenta que obtém o alinhamento múltiplo ótimo de múltiplas sequências. No entanto, essa ferramenta produz resultados exclusivamente através da linha de comando. Neste trabalho, propomos um web server para disponibilizar o acesso aos resultados com um navegador web. O web server foi planejado para ser implementado em múltiplas máquinas virtuais em um provedor de computação em nuvem e que podem ser aumentadas para se ajustar à demanda de trabalho.*

1. Introdução

A Bioinformática é uma área que envolve diferentes campos da ciência, tais como Ciência da Computação, Estatística, Química e Matemática. A comparação de sequências é um dos problemas fundamentais da Bioinformática e pode ser utilizado para a solução de diversos outros problemas. Na comparação de sequências são recebidas sequências biológicas e é produzido um alinhamento que permite a visualização das similaridades e diferenças entre elas.

O alinhamento de múltiplas sequências permite a comparação simultânea entre 3 ou mais sequências e o seu alinhamento permite a visualização de similaridades e diferenças entre elas. Existem dois tipos de métodos que alcançam os alinhamentos de sequências, que são globais e locais. Em alinhamentos globais ocorrem as análises das sequências inteiras. Já no alinhamento local são verificadas partes das sequências.

Os métodos heurísticos são algoritmos que têm como objetivo alcançar a redução do tempo de execução, resolvendo problemas complexos em menor tempo, mas que não garantem a obtenção do melhor resultado possível. Mesmo com a utilização de técnicas de otimização de algoritmos heurísticos, com a tentativa de reduzir os custos de processamentos computacionais, esses métodos não conseguem alcançar, por muitas das vezes, o alinhamento desejado. No entanto, seus resultados têm grandes chances de alcançar um alinhamento que seja muito próximo de uma conclusão a qual pode ser considerada ideal.

O PA-Star [Sundfeld et al. 2018] é um algoritmo proposto para a solução do alinhamentos global ótimo utilizando da técnica de busca de menor caminho entre dois nós em grafos e utiliza o algoritmo A* para garantir o resultado ótimo sem precisar explorar todos os nós do grafo. Esse algoritmo utiliza funções de *hash* sensíveis ao contexto para dividir o trabalho entre diversas threads.

Entretanto, o PA-Star possui uma interface somente por linha de comando dos sistemas operacionais Linux. Por isso, seu acesso pode ser limitado a usuários que não estejam familiarizados com a linha de comando ou o sistema operacional. Para resolver esse problema, neste trabalho propomos e implementamos o servidor web PA-Star-Web para que seja possível utilizar o PA-Star através de um navegador web.

2. Trabalhos Correlatos

Existem diversos trabalhos com a finalidade de implementar *web servers* em diversas áreas da Bioinformática [Passaro et al. 2020, Li et al. 2019, Havgaard et al. 2005].

OrtoVenn2 [Xu et al. 2019] é um web server de comparação de clusters ortólogos com um limite para comparações de até 12 sequências de espécies, além de disponibilizar um contêiner Docker para seus usuários executarem localmente sem que haja um limite máximo no número de comparações entre sequências.

[Lee et al. 2019] propuseram o DNAvisualization, uma aplicação Web para visualizar e analisar sequências de DNA através de gráficos bidimensionais. Para implementar, utiliza o recurso de computação sem servidor, utilizando as funções lambda da Amazon Web Service (AWS). Porém está limitado para alinhar um máximo de 30 sequências simultaneamente.

3. PA-Star-Web

O PA-Star-Web é implementado utilizando a linguagem de programação Python, utilizando o framework Django, que executa em múltiplas instâncias EC2 na Amazon AWS. A primeira instância é a principal. Nela, é executado o servidor HTTP que recebe requisições dos usuários e também devolve a eles os resultados produzidos. O segundo tipo de instância é a trabalhadora. Nela, são recebidas as requisições da instância principal, o programa PA-Star é executado e os resultados são retornados à instância principal. Pode-se utilizar instâncias na nuvem com diferentes capacidade de processamento.

A Figura 1 ilustra a nossa arquitetura com apenas uma instância trabalhadora e uma instância principal. Para gerenciar as tarefas e a fila de requisições é utilizada a ferramenta Celery, que permite enfileirar tarefas em filas Amazon Simple Queue Service (SQS) e o servidor web utilizado é o Ngnix.

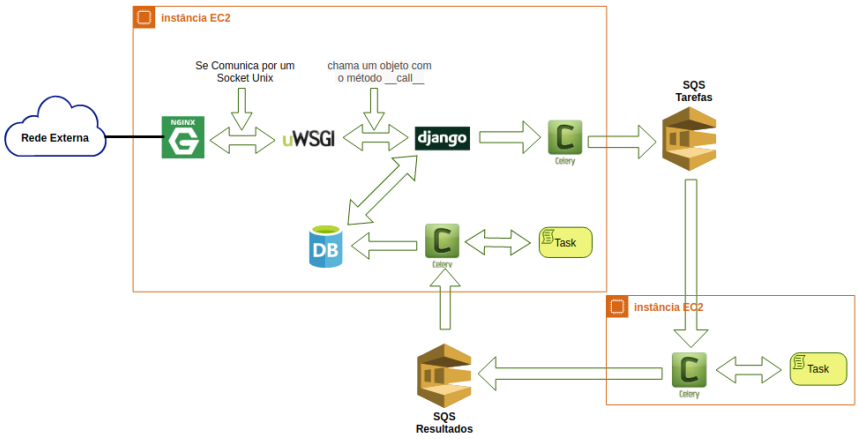

Figura 1. Arquitetura em nuvem proposta para implementação

O servidor web Ngnix é a porta de entrada para requisições de redes externas e é executado dentro de uma instância EC2 da Amazon AWS. Este servidor é capaz de servir de forma imediata arquivos estáticos como imagens, JavaScript e CSS. A comunicação entre o servidor web e o framework Django é realizada com o protocolo uWSGI. As solicitações de alinhamento de múltiplas sequências são enviadas com a Celery para uma fila Amazon SQS. Após enviar uma tarefa com êxito, um id único é gerado e enviado ao usuário para posterior visualização.

Em uma instância trabalhadora, as mensagens são consumidas da fila de tarefas também pela Celery. Múltiplas instâncias trabalhadoras podem aguardar simultaneamente na fila pois as mensagens são entregues para apenas uma instância. Ao receber uma tarefa, é realizada uma chamada de sistema para a execução do programa PA-Star. Após o alinhamento de múltiplas sequências ser concluído pela ferramenta PA-Star, os resultados são enviados em uma outra fila de resultados. Na instância principal, os resultados que chegam através desta fila são armazenados dentro de um banco de dados e então o resultado final fica disponível para apresentação ao usuário.

A Figura 2 apresenta a tela de resultado da requisição de uma análise de sequências. Nesta tela é mostrado o tempo de execução e o alinhamento gerado pelo algoritmo PA-Star, com algumas marcações e diferenciações em cores.

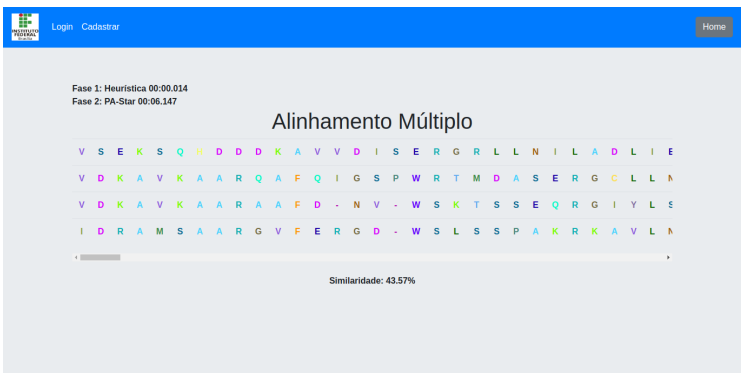

Figura 2. Tela de resultados

4. Conclusão e Trabalho Futuros

O uso de ferramentas de computação de alto desempenho pode ser limitado devido à complexidade de instalação e acesso limitado a recursos computacionais. A computação em nuvem pode diminuir essa complexidade, com o uso de recursos sob demanda e que podem ser provisionados automaticamente.

Neste trabalho, propusemos e implementamos um *web server* com uma arquitetura em nuvem para o PA-Star, criando uma aplicação web para o alinhamento múltiplo de sequências. Essa arquitetura é escalável e fracamente acoplada, de forma que os recursos podem ser adicionados para aumentar a capacidade computacional do sistema.

Como trabalhos futuros, desejamos automatizar o provisionamento de toda a infraestrutura utilizando *templates* Terraform ou AWS Cloudformation. Desejamos inserir um serviço de monitoramento e *scaling* automático para que as instâncias trabalhadoras possam ser adicionadas e terminadas automaticamente de acordo com a demanda de tarefas. Também desejamos avaliar a escalabilidade automática dessa solução. Finalmente, desejamos portar o serviço de execução do PA-Star para um ambiente com contêiners de forma que ele possa ser executado não apenas em instância EC2, mas em clusters compartilhados e gerenciados por provedores em nuvem, utilizando serviços de submissão de *jobs* como o AWS Batch.

Referências

Havgaard, J. H., Lyngsø, R. B., and Gorodkin, J. (2005). The foldalign web server for pairwise structural rna alignment and mutual motif search. *Nucleic acids research*, 33:W650–W653.

Lee, B. D., Timony, M. A., and Ruiz, P. (2019). DNAvisualization.org: a serverless web tool for DNA sequence visualization. *Nucleic Acids Research*, 47(W1):W20–W25.

Li, D., Hsu, S., Purushotham, D., Sears, R. L., and Wang, T. (2019). WashU Epigenome Browser update 2019. *Nucleic Acids Research*, 47(W1):W158–W165.

Passaro, M., Martinovic, M., Bevilacqua, V., Hershberg, E. A., Rossetti, G., Beliveau, B. J., Bonnal, R. J. P., and Pagani, M. (2020). OligoMinerApp: a web-server application for the design of genome-scale oligonucleotide in situ hybridization probes through the flexible OligoMiner environment. *Nucleic Acids Research*, 48(W1):W332–W339.

Sundfeld, D., Razzolini, C., Teodoro, G., Boukerche, A., and Melo, A. C. M. A. (2018). Pa-star: A disk-assisted parallel a-star strategy with locality-sensitive hash for multiple sequence alignment. *Journal of Parallel and Distributed Computing*, 112:154–165.

Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., Zhang, G., Gu, Y. Q., Coleman-Derr, D., Xia, Q., and Wang, Y. (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Research*, 47(W1):W52–W58.