CUDA-Sankoff-Web: Uma ferramenta web para cálculo do alinhamento secundário estrutural ótimo

Rodrigo Rocha Gomes[1], **Daniel Sundfeld**[1]

[1]Instituto de Educação, Ciência e Tecnologia de Brasília – *Campus* Brasília
Setor de Grandes Áreas Norte 610 – Asa Norte – 70830-450 – Brasília – DF – Brasil

rodrigo.gomes@estudante.ifb.edu.br, daniel.sundfeld@ifb.edu.br

Resumo. *O alinhamento de sequências é uma das operações mais importantes para a Bioinformática. Ao se considerar sequências de RNA, não basta considerar apenas a sequência em si, mas também devem ser consideradas as estruturas formadas por elas. O alinhamento ótimo de sequências possui alta complexidade computacional e pode demorar muito tempo. Recentemente, o CUDA-Sankoff foi proposto para obter o alinhamento ótimo secundário de sequências de RNA em um tempo razoável usando unidades de processamento gráfico (GPUs). No entanto, essa ferramenta não possui uma interface amigável. Neste trabalho, propomos criar um web server que provê uma interface gráfica e facilita a execução da ferramenta CUDA-Sankoff.*

1. Introdução

Bioinformática é um campo interdisciplinar com foco principal nas subáreas da Biologia [Mount 2001]. Com a evolução da tecnologia computacional e das pesquisas biológicas, tornou-se necessário o auxílio de computadores para realizar procedimentos específicos da área, como um algoritmo que compara duas ou mais partes do genoma de uma espécie a fim de encontrar padrões evolutivos, por exemplo. A partir disso, diversos algoritmos para elucidar problemas similares foram desenvolvidos.

A comparação de sequências é uma das operações fundamentais da Bioinformática. Dadas suas sequências, determinar quais são as semelhanças e diferenças é muito importante, pois permite traçar uma história evolutiva da espécie em estudo. Ao comparar sequências, métodos tradicionais realizam apenas a comparação de nucleotídeos, como por exemplo o algoritmo de Smith-Waterman [Smith et al. 1981]. Essa comparação é conhecida como comparação primária. O RNA é conhecido por poder dobrar sobre si mesmo e formar uma estrutura. Muitas vezes a função biológica é desempenhada pela estrutura, sendo assim algoritmos de comparação tradicionais podem não detectar uma semelhança entre as sequências.

2. Estrutura Secundária de RNA

Os nucleotídeos de uma sequência de RNA podem interagir entre si e formar estruturas complexas. Há diversas alternativas para representar uma estrutura secundária de RNA. A estrutura secundária é uma representação simplificada da estrutura real de RNA. A estrutura secundária possui muitas informações relevantes e comumente é utilizada como base para o estudo das estruturas mais complexas [Gorodkin and Ruzzo 2014].

As regiões mais comuns em uma estrutura de RNA são: a) Região Stem (Talo): Região que possui um ou mais nucleotídeos pareados em ordem; b) Região Loop (Laço) Quando dois nucleotídeos estão pareados e há nucleotídeos encadeados não pareados; c) Região Hairpin (Grampo): Ocorre quando a região steam está adjacente a região loop; d) External Loop (Laço Externo): Região onde não ocorre pareamento entre as bases e se encontra no final da sequência de RNA; e) Internal Loop (Laço Interno) : Região adjacente a duas regiões *steams*, ou um único par de bases, contendo nucleotídeos não pareados entre eles; f) Multibranch (Junção): Também chamada de multibranch-loop. Região que contém nucleotídeos não-pareados e que combina outras sub-regiões em estruturas mais complexas.

3. CUDA-Sankoff

O Algoritmo de Sankoff possui uma alta complexidade computacional $(O(n^6))$ e pode demorar muito tempo para obter resultados. Para resolver esse problema, o CUDA-Sankoff [Sundfeld et al. 2020] propõe utilizar unidades de processamento gráfico (GPU) com a arquitetura CUDA para acelerar o processamento do algoritmo.

Para a paralelização do algoritmo, o CUDA-Sankoff propôs a criação de uma matriz de programação dinâmica de 4 dimensões (4D-DP). Esta matriz pode ser visualizada como uma matriz externa (ME) 2D, onde cada célula é uma outra matriz, chamada de matriz interna (MI), também 2D. A equação de recorrência do algoritmo mostra que, de acordo com as dependências de dados, as anti-diagonais da matriz externa e as anti-diagonais das matrizes internas podem ser calculados em paralelo, em um padrão de *wavefront*. Nos resultados obtidos, comparado com uma versão baseada em CPU utilizando 32 núcleos, o CUDA-Sankoff é capaz de atingir um *speedup* de 7,81x utilizando uma NVidia Tesla P100, reduzindo o tempo total de execução de 6 horas e 18 minutos para 48 minutos e 20 segundo.

4. Trabalhos Correlatos

Com o crescente acesso a dados do genoma e sua imensa quantidade de informação biológica, [Lee et al. 2019] apresentam uma plataforma web para visualização de dados biológicos. Através de técnicas de computação paralela e recuperação seletiva de dados, o site é hábil para a visualização de sequências simultâneas de DNA.

Comparar sequências de RNA é uma técnica utilizada por pesquisadores no campo de estudo de Genomas a fim de localizar similaridades entre duas espécies distintas. Mas essa ferramenta exige muito poder de processamento da máquina que o utiliza. [Mallawaarachchi et al. 2018] utiliza a arquitetura CUDA (Compute Unified Device Architecture) para processamento simultâneo e paralelo entre o CPUs e GPUs. Além disso, utiliza cluster a fim de otimizar o tempo de processamento entre as sequências.

O trabalho [Angiuoli et al. 2011] disponibiliza uma ferramenta desktop portátil no formato de uma Máquina Virtual. Dessa forma, atua em todos os sistemas operacionais com suporte a virtualização. A ferramenta oferece diversos pipelines para genômica microbiana, incluindo 16S, genoma inteiro ou análise de sequências de metagenoma. O CloVR é utilizado em um computador pessoal, utiliza recursos locais e requer mínima instalação. Além disso, há possibilidade de utilizar recursos de nuvem para aprimorar o desempenho do processamento em larga escala.

5. Trabalho Proposto

Neste trabalho, propomos o CUDA-Sankoff-Web, uma ferramenta web que possibilita o cálculo do alinhamento estrutural ótimo na nuvem e facilita a visualização dos dados. O trabalho realizado foi desenvolvido utilizando o framework Django e foi projetado para executar em uma única máquina virtual. O CUDA-Sankoff-Web disponibiliza uma página web onde o usuário pode realizar o upload de um arquivo FASTA ou preencher um formulário com as sequências.

A Figura 1 ilustra o funcionamento do CUDA-Sankoff-Web. Ao receber os dados do usuário, o CUDA-Sankoff-Web realiza uma chamada de sistema para executar o CUDA-Sankoff. Após o término da execução, a saída do programa CUDA-Sankoff é processada para mostrar cores nos nucleotídeos iguais, facilitando a visualização. Além disso, o CUDA-Sankoff-Web também calcula o "GC-Content" das sequências, que é a proporção entre Guanina e Citosina, informação relevante ao se analisar sequências de RNA. Além disso, também é produzida uma representação gráfica da estrutura secundária do RNA utilizando a ferramenta RNAplot disponível na biblioteca ViennaRNA [Lorenz et al. 2011]. Todas essas informações são colocadas em uma página de resultados.

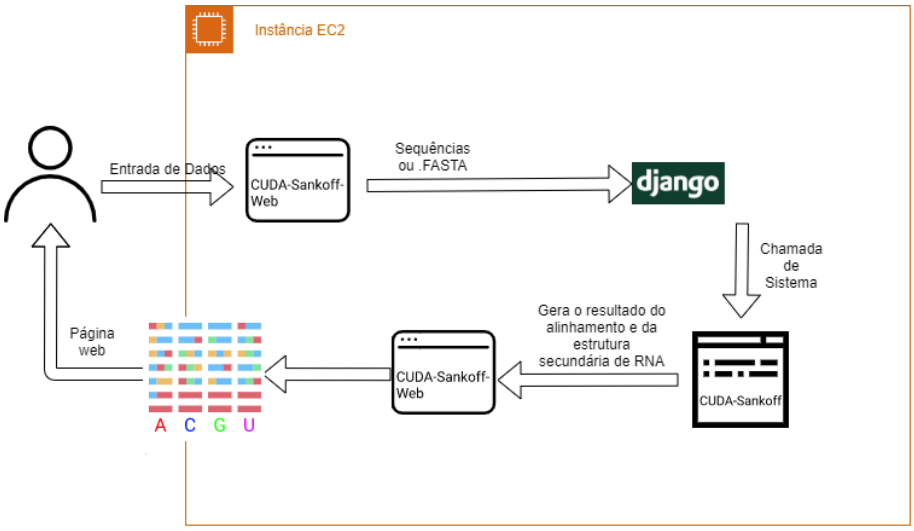

Figura 1. Fluxo de funcionamento do CUDA-Sankoff-Web

O CUDA-Sankoff-Web possui um instalador, de forma que o usuário possa criar instâncias privadas. Para isso, basta provisionar uma instância de máquina virtual Linux que atende às suas necessidades, baixar o código-fonte da nossa ferramenta através de um repositório de código-fonte Git e executar um script de instalação, que irá resolver todas as dependências necessárias e iniciar o serviço. O instalador do CUDA-Sankoff-Web está adaptado para utilizar a Elastic Compute Cloud (Amazon EC2) da Amazon Web Services (AWS).

O CUDA-Sankoff exige grande poder computacional. Na maioria das instâncias esse processamento será realizado na CPU. Mas se a instância criada tiver uma GPU e o ambiente de desenvolvimento CUDA, será utilizada a placa aceleradora para o processamento e pode-se obter os resultados com um tempo muito menor de execução.

6. Conclusão e Trabalhos Futuros

O uso de ferramentas de computação de alto desempenho pode ser limitado aos cientistas de áreas externas à computação pois normalmente exigem hardware especializado e conhecimentos técnicos especializados para instalação, configuração e manutenção do sistema. Neste trabalho, foi proposto um sistema web para que os usuários utilizem serviços tradicionais de computação em nuvem para ter acesso a uma ferramenta de computação de alta performance. Esse sistema visa aumentar o alcance e uso da ferramenta pela comunidade científica, além de permitir uma interface mais amigável, que inclui informações gráficas e mais completas para a análise das sequências.

Além disso, todos os componentes da ferramenta executam em uma única instância EC2. Como trabalho futuro, propomos a otimização dos recursos utilizados pela ferramenta dividindo as múltiplas partes do sistema em diversas instâncias na nuvem, criadas sob demanda. Desta forma, atividades simples como receber arquivos e validar a entrada, podem ser executadas em instâncias baratas com pouco poder de processamento e memória RAM. Ao receber uma tarefa, é criada uma instância sob demanda para processamento. Esta instância é mais cara por possuir mais núcleos, RAM e GPU, mas finaliza ao terminar a tarefa e com isso os custos podem ser otimizados.

Referências

Angiuoli, S. V., Matalka, M., Gussman, A., Galens, K., Vangala, M., Riley, D. R., Arze, C., White, J. R., White, O., and Fricke, W. F. (2011). Clovr: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC bioinformatics*, 12(1):356.

Gorodkin, J. and Ruzzo, W. L. (2014). *RNA sequence, structure, and function: computational and Bioinformatic methods*. Springer.

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.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):26.

Mallawaarachchi, V., Wickramarachchi, A., Welivita, A., Perera, I., and Meedeniya, D. (2018). Efficient bioinformatics computations through gpu accelerated web services. In *Proceedings of the 2018 2nd International Conference on Algorithms, Computing and Systems*, pages 94–98. ACM.

Mount, D. W. (2001). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 1st edition.

Smith, T. F., Waterman, M. S., et al. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

Sundfeld, D., Teodoro, G., Havgaard, J. H., Gorodkin, J., and Melo, A. C. (2020). Using GPU to accelerate the pairwise structural RNA alignment with base pair probabilities. *Concurrency and Computation: Practice and Experience*, 32(10):e5468.