Comparação de Sequências Biológicas em Cluster de GPUs na Nuvem

Walisson P. Sousa[1], Filipe M. Soares[2], Alba C. M. A. Melo[2], Cristiana Bentes[1], Maria Clicia S. de Castro[1]

[1]Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro – RJ – Brasil

[2]Universidade de Brasília (UNB)
Brasília – DF – Brasil

walisson.sousa@pos.ime.uerj.br, 150051409@aluno.unb.br, alves@unb.br
cris@eng.uerj.br,clicia@ime.uerj.br

Abstract. *Genomic sequence comparison is a process used to find structural alterations in genes or proteins of living organisms. It is simple, computationally expensive, and needs algorithms to reduce the processing time. This paper assesses the monetary cost and runtime when comparing genomic sequences in cloud instances with GPUs. Experimental results show that comparing longer sequences is advantageous.*

Resumo. *A comparação de sequências genômicas é um processo utilizado para encontrar alterações estruturais em genes ou proteínas de organismos vivos. Ela é simples, computacionalmente custosa e necessita de algoritmos que reduzam seu o tempo de processamento. Este trabalho avalia o custo monetário e tempo de execução da comparação de sequências genômicas em instâncias na nuvem com GPUs. Resultados experimentais mostram que comparar sequências maiores nos clusters é vantajoso.*

1. Introdução

A comparação de sequências biológicas é uma das operações básicas mais importantes da Bioinformática. Ela identifica alterações estruturais em genes ou proteínas de organismos vivos [Batzoglou 2005]. Nesta operação são selecionadas duas sequências de entrada utilizadas para criar uma matriz de programação dinâmica e calcular uma pontuação (*score*), que indica o nível de similaridade entre as sequências.

Existem dois tipos principais de alinhamentos: o global (melhor correspondência entre as sequências) e o local (melhor correspondência entre partes da sequência). Um algoritmo local que compara duas sequências de forma exata é o Smith-Waterman (SW) [Smith et al. 1981], que é dinâmico e baseado no problema da Maior Subsequência Comum (LCS). Ele calcula uma matriz $[(n+1)$ X $(m+1)]$ de similaridade com complexidade quadrática ($O(mn)$). As Soluções com métodos heurísticos foram desenvolvidas para obter a comparação mais rápida, mas não garantem a solução ótima.

Outra abordagem para redução do tempo implementa versões paralelas do SW em Unidades Gráficas de Processamento (GPUs). O MASA-CUDAlign é uma ferramenta que utiliza algoritmos exatos, variantes do SW, para realizar a comparação de sequências

biológicas. Ela possui versões para diversos *clusters* físicos com bons resultados (82,82 TCUPS) [Figueirêdo Júnior 2021], mas não exploram *clusters* em nuvem. Plataformas da Amazon (AWS), Microsoft Azure, Google Cloud e Alibaba Cloud permitem o uso de recursos computacionais sob demanda com menor custo, sendo ideal para tarefas de alto desempenho sem aquisição de infraestrutura específica.

O ParallelCluster é uma ferramenta baseada nos recursos do Elastic Computing Cloud (EC2) e de outros serviços da AWS. Sua estrutura é composta por um nó mestre e nenhum ou mais nós computacionais, cujos principais recursos são: Instâncias EC2, Elastic Block Store (EBS) e Virtual Private Cloud (VPC). Neste contexto, os serviços em nuvem são uma alternativa atrativa, porque existe uma vasta oferta recursos computacionais no mercado. Este trabalho realiza a avaliação do custo monetário e tempo de execução considerando o ParallelCluster na comparação de sequências biológicas em *clusters* de GPUs, excutando a ferramenta MASA-CUDAlign.

Este trabalho está organizado nas quatro seções. A Seção 2 aborda as comparações de sequências biológicas. A Seção 3 apresenta o ambiente, experimentos e os resultados obtidos. As conclusões e propostas futuras são descritas na Seção 4.

2. Comparação de Sequências Biológicas

Comparar sequências genéticas de forma exata é uma tarefa trivial que exige alto poder computacional dependendo do comprimento das cadeias. As sequências que representam cromossomos do DNA humano possuem entre 50 a 300 milhões caracteres.

A Figura 1 mostra uma operação de comparação entre as cadeias $S_0 = AGTTCCGGAGG$ e $S_1 = ACTTCCAGA$. A semelhança entre S_0 e S_1 é aferida pela pontuação, calculada de acordo com a correlação entre seus caracteres. Os pontos são computados a cada par de caracteres na mesma posição das sequências. Se os caracteres de S_0 e S_1 forem iguais, assume-se que existe uma correlação (*match*) e é atribuído um valor positivo (+1 no exemplo). Se os caracteres forem diferentes, existe uma incompatibilidade (*mismatch*) e é atribuído um valor negativo. Ou ainda, se um dos caracteres for uma lacuna (*gap*) é atribuído um valor negativo.

A	C	T	T	C	C	-	-	A	G	A
A	G	T	T	C	C	G	G	A	G	G
+1	-1	+1	+1	+1	+1	-2	-2	+1	+1	-1

$\sum = 1$

Figura 1. Alinhamento entre S_0 e S_1 com *score* 1 [Sandes and Melo 2010].

A variação do SW utilizada para comparar cada posição de S_0 com a de S_1, resulta numa matriz de similaridade (Figura 2 (esquerda). O preenchimento da matriz é feito de acordo com o elemento da linha anterior, da coluna anterior e da diagonal superior esquerda, utilizando os valores de *match*, *mismatch* e *gap*. O MASA-CUDAlign calcula os *gaps* usando o modelo *affine gap* [Gotoh 1982], na versão modificada por [Myers and Miller 1988] para trabalhar com espaço linear através de divisão e conquista, que atribui uma penalidade maior para o primeiro *gap* na incidência de *gaps* consecutivos.

Essa versão paralela do SW é baseada em CUDA[1]. A matriz de similaridade é

[1] `https://developer.nvidia.com/cuda-zone`

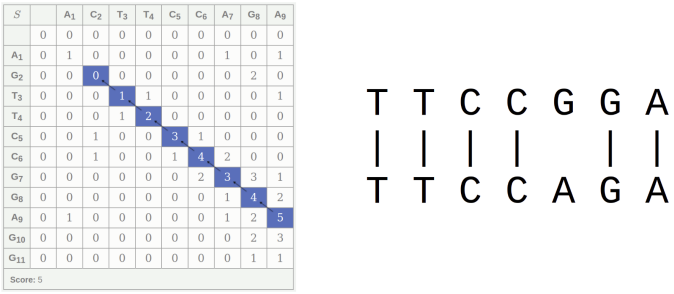


Figura 2. Matriz de similaridade entre S_0 e S_1 e o seu alinhamento

dividida em blocos, executados paralelamente em diagonal, devido à dependência da pontuação da comparação do caractere anterior.

3. Ambiente e Resultados Obtidos

A Tabela 1 mostra números de acesso, nome dos organismos, tamanho das cadeias e *score* das sequências utilizadas nos experimetos e obtidas do Centro Nacional de Informações sobre Biotecnologia [NCBI 2021].

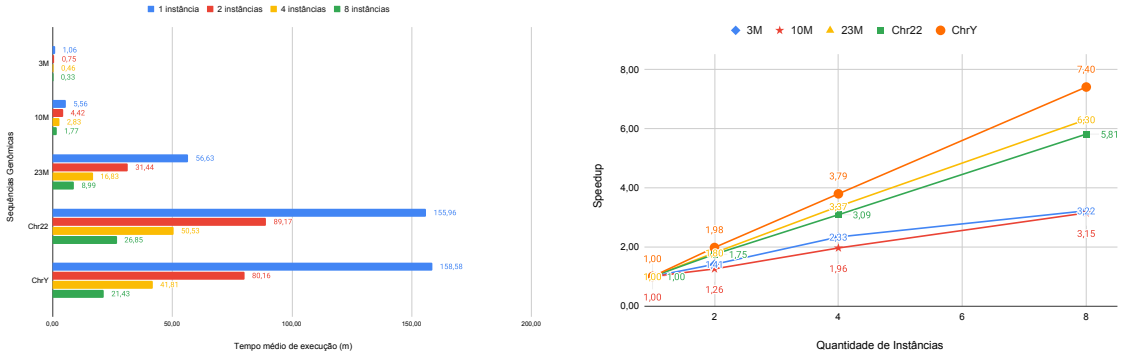
Tabela 1. Sequências genômicas utilizadas nos experimentos.

Seq.	Acesso	Nome	Tamanho	Score
3M	BA000035.2 BX927147.1	*Corynebacterium efficiens YS-314* *Corynebacterium glutamicum ATCC 13032*	3.147.090 3.282.708	4226
10M	NC_014318.1 NC_017186.1	*Amycolatopsis mediterranei U32 chromosome* *Amycolatopsis mediterranei S699 chromosome*	10.236.715 10.236.779	10235188
23M	NT_033779.4 NT_037436.3	*Drosophila melanogaster chromosome 2L* *Drosophila melanogaster chromosome 3L*	23.011.544 24.543.557	9063
Chr22	NC_000022.11 NC_006489.4	*Homo sapiens chromosome 22* *Pan troglodytes chromosome 22*	50.818.468 37.823.149	20426645
ChrY	NC_000024.10 NC_006492.4	*Homo sapiens chromosome Y* *Pan troglodytes chromosome Y*	57.227.415 26.350.515	1448250

Os testes foram executados no ParallelCluster da AWS, na zona *us-east-1* com o EBS para armazenamento. As comparações utilizaram a versão MASA-CUDAlign static-multibp [de Figueiredo Júnior et al. 2021]. A construção do *cluster* considerou a instância g4dn.xlarge, utilizada em trabalho anterior [Brum et al. 2021], e que apresentou boa relação custo-benefício, com menos de um dólar por hora de utilização. A instância é composta de processador Intel Xeon 24C 2.5 GHz, 16GB de RAM e GPU Nvidia T4.

Foram criados quatro cenários de execução: com 1 instância e *clusters* com 2, 4 e 8 instâncias, além do nó mestre. Cada cenário foi executado 5 vezes e apresentou desvio padrão abaixo de 5%. O MASA-CUDALign foi executado em todas as comparações mostradas na Tabela 1. A Figura 3(a) apresenta o tempo médio nos 4 cenários. Houve uma redução no tempo ao dobrar o número de instâncias do *cluster*, sendo mais evidente nas sequências maiores.

A métrica de *speedup* foi utilizada para avaliar o desempenho da comparação das sequências em cada um dos cenários, como mostra a Figura 3(b). O *speedup* das comparações das sequências maiores (23M, Chr22, ChrY) é maior em relação às sequências menores. O ganho ao dobrar o número de instâncias do *cluster*, apesar de existir, é menos evidente nas comparações das sequências 3M e 10M.



(a) (b)

Figura 3. Tempo médio de execução e speedup nos cenários com 1, 2, 4 e 8 instâncias

A estimativa do custo é dada pela Equação 1 considerando três fatores:(i) n_{hosts}: quantidade de instâncias, que compõem o *cluster*; (ii) t_{medio}: tempo médio de execução, em s; e (iii) $Preco_{instancia}$: Preço da instância por hora de utilização em dólares.

$$Custo_{estimado} = \frac{n_{hosts} \cdot t_{medio} \cdot Preco_{instancia}}{3600} \quad (1)$$

A Figura 4 apresenta os custos de execução estimados para cada um dos cenários. É importante ressaltar que o tempo médio de execução é referente à execução de todas as comparações, uma única vez, e que o preço varia de acordo com o tipo de instância. Neste caso, a instância custou $0,526$ dólares e o custo foi convertido para reais (R$) com o valor médio do dólar no período dos testes (R$ 5,39).

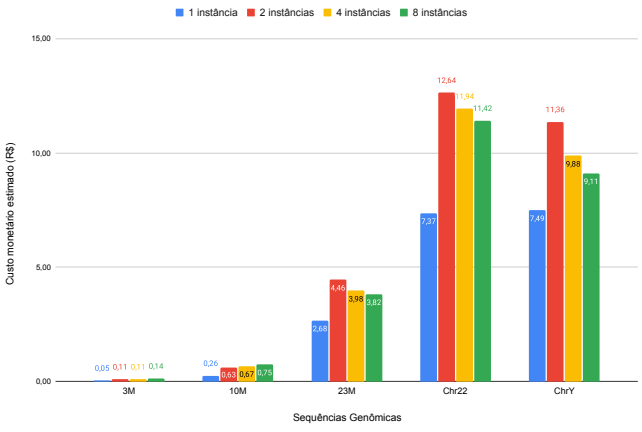


Figura 4. Custo monetário estimado (R$)

O menor custo monerátio estimado foi obtido no cenário com uma instância. Os *clusters* com 2 instâncias apresentam maior custo, exceto para a sequência de 10M. É possível notar que nos *clusters* com 4 e 8 instâncias o custo estimado é mais baixo. Isso ocorre devido ao nó mestre, que tem seu custo diluído quando há aumento na quantidade de instâncias.

4. Considerações Finais

Este trabalho avaliou a ferramenta MASA-CUDAlign em *cluster* com GPU da AWS, considerando custo e tempo de execução. Resultados mostram que aumentar o tamanho do *cluster* beneficia as sequências maiores. No futuro está previsto aumentar a quantidade de instâncias com GPU, usar múltiplas GPUs por instância, e usar outras abordagens para prever o tamanho do *cluster* e os custos associados.

Referências

Batzoglou, S. (2005). The many faces of sequence alignment. *Briefings in bioinformatics*, 6(1):6–22.

Brum, R. C., Sousa, W. P., Melo, A. C., Bentes, C., Castro, M. C. S. d., and Drummond, L. M. d. A. (2021). A fault tolerant and deadline constrained sequence alignment application on cloud-based spot gpu instances. In *European Conference on Parallel Processing*, pages 317–333. Springer.

de Figueiredo Júnior, M. A. C., Navarro, J. P., de Oliveira Sandes, E. F., Teodoro, G., and Melo, A. C. M. (2021). Parallel fine-grained comparison of long dna sequences in homogeneous and heterogeneous gpu platforms with pruning. *IEEE Transactions on Parallel and Distributed Systems*.

Figueirêdo Júnior, M. A. C. d. (2021). Comparação paralela de sequências biológicas em múltiplas gpus com descarte de blocos e estratégias de distribuição de carga.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, 162(3):705–708.

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

NCBI (2021). National Center for Biotechnological Information. `https://www.ncbi.nlm.nih.gov/`.

Sandes, E. and Melo, A. (2010). Cudalign: using gpu to accelerate the comparison of megabase genomic sequences. In *Proc. of the 15th ACM SIGPLAN*, pages 137–146.

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