Análise de Desempenho do PA-Star2 no SDumont e sua Aplicação em um *Workflow* Científico

Kelen Souza[1,2], **Rafael Terra**[1], **Carla Osthoff**[1], **Kary Ocaña**[1], **Hiago Rocha**[1]

[1]Laboratório Nacional de Computação Científica (LNCC)
[2]Faculdade de Educação Tecnológica do Rio de Janeiro (FAETERJ)
Petrópolis – RJ, Brasil

{kelenbs, rafaelst, osthoff, karyann, mayk}@lncc.br

Abstract. *In this study, the problem of multiple sequence alignment (MSA) was addressed, limited by the small number of input sequences in an exact algorithm software for CPUs: PA-Star. The performance of PA-Star1 and PA-Star2 (the current version) was compared through tests carried out on the Santos Dumont (SDumont) supercomputer. The results showed significant reductions in alignment execution time with the most recent version. Therefore, PA-Star2 was integrated into a scientific workflow, aiming to expand the software's usability. This application demonstrated effectiveness and contributed to a significant reduction in processing time when handling larger sets of sequences.*

Resumo. *Neste estudo, foi abordado o problema do alinhamento múltiplo de sequências (AMS), limitado pelo pequeno número de sequências de entrada em um software de algoritmo exato para CPUs: o PA-Star. Comparou-se o desempenho das versões PA-Star1 e PA-Star2 (atual) em testes realizados no supercomputador Santos Dumont (SDumont). Os resultados mostraram reduções expressivas no tempo de execução dos alinhamentos na versão mais recente. Portanto, o PA-Star2 foi integrado a um workflow científico, visando ampliar o uso do software. Essa aplicação demonstrou eficácia e contribuiu para uma redução significativa no tempo de processamento ao lidar com conjuntos maiores de sequências.*

1. Introdução

Na Bioinformática, o alinhamento múltiplo de sequências (AMS) é uma etapa crítica na avaliação de sequências biológicas, pois permite identificar regiões de similaridade entre DNA, RNA ou proteínas, revelando relações funcionais, estruturais e evolutivas. No entanto, o AMS é classificado como um problema NP-difícil, que impõe grande desafio computacional e exige grande espaço de memória [Sundfeld et al. 2018]. Apesar dessas limitações, encontrar o alinhamento ótimo ainda é o objetivo, que será referência de qualidade frente a outros métodos, que embora mais escaláveis, podem introduzir aproximações e perda de precisão.

Este estudo investiga portanto o uso do PA-Star, programa desenvolvido por pesquisadores da UnB, colaboradores nesta pesquisa. É uma versão paralela baseada no algoritmo exato A-star (A*), desenvolvido para resolver o problema de AMS utilizando CPUs. O PA-Star2, sua versão atual, apresenta melhorias de desempenho, com ênfase

na otimização da divisão de tarefas em máquinas com processadores assimétricos (*Asymmetric Multicore Processors* – AMPs) [Sundfeld et al. 2025]. Embora existam diversas ferramentas de algoritmo exato, a limitação enfrentada quanto ao número de sequências de entrada se mantém, mesmo com o uso de HPC.

Os testes foram realizados no supercomputador Santos Dumont (SDumont)[1]. Nesta pesquisa foram comparadas as versões do programa, PA-Star1 (anterior) e PA-Star2, identificando a versão 2 como a de melhor desempenho, sendo, portanto, escolhida com base nos resultados para integrar o *workflow* científico de bioinformática automatizado com PyCOMPSs no trabalho de [Terra et al. 2025]. Um *workflow* científico é um processo de etapas computacionais com ferramentas e dados para analisar, processar e interpretar informações biológicas de forma reprodutível. Foi utilizado o PyCOMPSs, que é a interface Python do COMPSs, um modelo de programação e ambiente de execução, que oferece uma interface sequencial, mas explora paralelismo em tempo de execução [2].

Esse processo representado na figura abaixo (simplificado de [Terra et al. 2025]) inicia com sequências em um arquivo multi-FASTA como entrada, a ferramenta MASA-OpenMP seleciona cada sequência e executa o alinhamento par-a-par, atribuindo um escore total a cada alinhamento. Com base nesse escore, é possível filtrar as sequências em código Python. As sequências selecionadas são então reagrupadas em um novo arquivo e executadas no PA-Star2, produzindo como resultado final um AMS completo e porcentagem de similaridade.

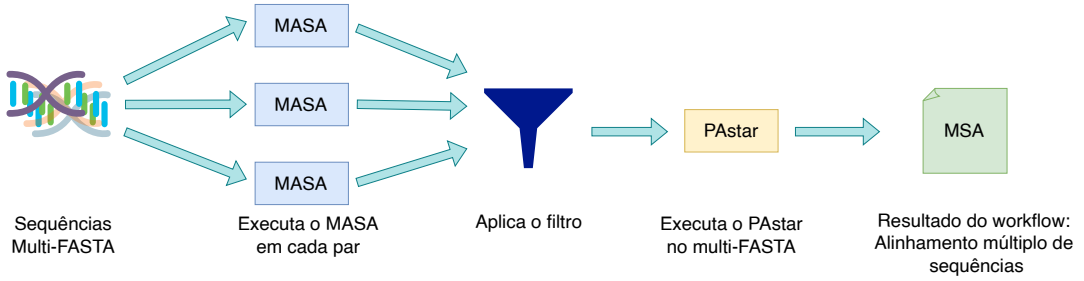

Figura 1. Diagrama básico do *workflow* científico desenvolvido com PyCOMPSs.

2. Metodologia

2.1. Ambiente Computacional

O supercomputador Santos Dumont (SDumont), do Laboratório Nacional de Computação Científica (LNCC), é composto por 376 nós computacionais, totalizando 18.048 núcleos de CPU e 376 GPUs. A máquina utilizada nesta pesquisa, Sequana, apresenta capacidade teórica total de 4,0 Pflops (RPeak). Foram utilizados os nós de CPU, cada um equipado com dois processadores Intel Xeon Cascade Lake Gold 6252, totalizando 48 núcleos de processamento e 384 GB de memória RAM. Para testes com maior demanda de memória, foram utilizados os nós com 768 GB de memória RAM[3].

[1] https://sdumont.lncc.br/
[2] https://pypi.org/project/pycompss/
[3] https://github.com/lncc-sered/manual-sdumont/wiki/

2.2. Dados de Entrada

Comparação PA-Star1 e PA-Star2. Foram usados arquivos multiFASTA de proteínas do banco de dados BAliBASE[4]. Os testes iniciaram com um conjunto de arquivos aqui nomeados como iniciais, sendo eles glg, 1sbp, 1aboA, 1ac5 e um segundo grupo de arquivos, classificados como intermediários: gal4, 1gpb, arp, 1sesA, 2myr, 2cba, 1hvA, 2ack e actin.

Análise do *Workflow*. Foram utilizadas sequências biológicas do vírus da dengue (DENV) do banco de dados GenBank[5]. Essas sequências foram organizadas em um *dataset* em diferentes versões: 9, 18, 28 e 38 sequências.

3. Resultados

Comparação entre PA-Star1 e PA-Star2. Os testes realizados com o PA-Star2 permitiram uma análise do comportamento da ferramenta e desempenho comparando-a com o PA-Star1. Os testes repetidos com os mesmos arquivos de proteínas (aminoácidos), nas mesmas configurações, revelaram uma redução expressiva no tempo de execução do alinhamento de, na maioria dos casos, 50% ou mais, como no caso do gal4 que chega a 68.6%. Com base nos resultados obtidos a versão 2 foi escolhida a fim de integrar um *workflow* científico, proposto em [Terra et al. 2025].

Tabela 1. Dados e comparação de resultados dos testes no PA-Star1 e PA-Star2

Arquivos (FASTA)	Tamanho do Arquivo	Nº de Seqs	Menor Seq	Maior Seq	Similaridade	Tempo (PA-Star1)	Tempo (PA-Star2)	RAM (PA-Star1)	RAM (PA-Star2)
glg	2.4K	5	438	486	26.80%	01m:58s	01m:01s	6,25 GB	5,40 GB
1sbp	1.3K	5	224	263	12.36%	01m:31s	00m:49s	4,28 GB	3,30 GB
1aboA	335	5	49	80	28.75%	01m:03s	00m:49s	5,08 GB	2,94 GB
1ac5	1.8K	4	421	483	25.10%	00m:59s	00m:26s	2,99 GB	1,69 MB
gal4	1.9K	5	335	395	15.80%	03h:43m:59s	01h:10m:18s	297,77 GB	249,66 GB
1gpb	4.0K	5	796	828	42.60%	01h:01m:12s	21m:50s	132,82 GB	110,15 GB
arp	2.1K	5	380	418	24.16%	21m:40s	08m:22s	48,78 GB	39,81 GB
1sesA	2.2K	5	417	442	29.87%	12m:41s	05m:20s	32,27 GB	25,72 GB
2myr	1.7K	4	340	474	14.94%	05m:17s	02m:01s	19,02 GB	13,16 GB
2cba	1.3K	5	237	259	22.15%	04m:07s	01m:57s	12,77 GB	8,71 GB
1hvA	928	5	136	199	14.07%	03m:35s	01m:44s	10,38 GB	8,53 GB
2ack	2.4K	5	452	482	18.82%	02m:14s	01m:09s	7,90 GB	6,60 GB
actin	2.0K	5	379	395	40.25%	01m:49s	00m:57s	6,69 GB	3,84 GB

Análise da Execução do *Workflow*. Essa seção avalia o *workflow* proposto por [Terra et al. 2025] com a integração do PA-Star2. Para isso, foram utilizados, como *baseline*, testes que representam a forma tradicional pela qual um biólogo executaria manualmente uma filtragem de sequências seguida de um AMS. Essa abordagem seria feita em série e sem paralelismo ou gerenciamento das tarefas. Entretanto, com a aplicação da ferramenta MASA-OpenMP como um filtro e execução posterior no PA-Star2 das sequências selecionadas, através de um *workflow* automatizado com PyCOMPSs[6], pode-se observar uma redução significativa no tempo de processamento, principalmente em

[4]https://www.lbgi.fr/balibase/

[5]https://www.ncbi.nlm.nih.gov/genbank/

[6]https://github.com/kelen-souza/Workflow-MASAOpenMP-PAStar

casos que o conjunto inicial de sequências de trabalho se mostra custoso demais para as ferramentas.

Os resultados mostraram que, embora o PA-Star2 aceite um número limitado de entradas diretas (no máximo 9 sequências), a integração com o *workflow* permite a seleção de sequências desejadas em um grupo considerável de sequências (como no caso da versão de 48 sequências). Reduzindo portanto o número de sequências para 5 por conjunto, sendo possível o AMS. Isso confirma que a ferramenta, embora restrita em termos de entrada direta, pode ser adaptada com sucesso para fluxos de trabalho automatizados e otimizados, desde que o volume de dados seja processado na etapa de filtragem do *workflow*.

Tabela 2. Comparação do tempo de processamento entre uma execução normal em série (*Baseline*) e em um *Workflow* com PyCOMPSs

Nº de sequências	Tempo (*Baseline*)	Tempo (*Workflow* com PyCOMPSs)
9	19,6s	38,96s
18	94,2s	50,98s
28	126,8s	32,43s
38	194,2s	37,26s

4. Conclusão

Os testes com o PA-Star2 evidenciaram reduções expressivas, de pelo menos metade ou mais, no tempo de execução do alinhamento. Com a atualização do programa, a melhor divisão de trabalho entre diferentes processadores aproveita melhor também as CPUs disponíveis. A integração do PA-Star2 a um *workflow* automatizado reduziu o número de sequências a serem alinhadas para 5 por conjunto, tornando viável o AMS mesmo com conjuntos significativos de sequências. Assim, amplia-se o potencial de uso da ferramenta em diferentes tipos de estudos, como análises filogenéticas e outras computacionalmente mais exigentes.

Referências

[Sundfeld et al. 2018] 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.

[Sundfeld et al. 2025] Sundfeld, D., Teodoro, G., and Melo, A. C. M. A. (2025). PA-Star2: Fast optimal multiple sequence alignment for asymmetric multicore processors. In *33rd Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP 2025)*, pages 146–153, Torino, Italy.

[Terra et al. 2025] Terra, R., Souza, K., Rocha, H., Osthoff, C., Carvalho, D., and Ocana, K. (2025). Workflow para alinhamento exato de sequencias em sistemas de processamento de alto desempenho. In *Anais do XXVI Simposio em Sistemas Computacionais de Alto Desempenho (SSCAD 2025)*, Bonito, MS, Brasil. Sociedade Brasileira de Computação, SBC.