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) – RJ/Brasil

[2]Faculdade de Educação Tecnológica do Rio de Janeiro (FAETERJ - Petrópolis) – RJ/Brasil

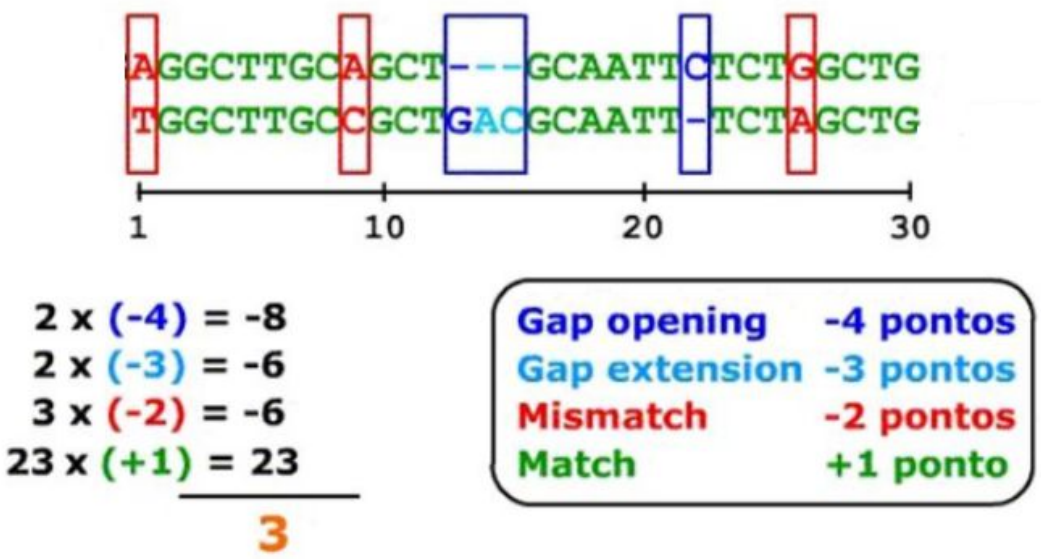
{kelenbs, rafaelst, osthoff, karyann, mayk}@lncc.br

FÓRUM DE INICIAÇÃO CIENTÍFICA - X ESCOLA REGIONAL DE ALTO DESEMPENHO DA REGIÃO SUDESTE (ERAD-SE 2025)

Sumário

Motivação

- O **alinhamento múltiplo de sequências (AMS)** é uma **técnica fundamental na bioinformática** por revelar **relações funcionais e evolutivas entre sequências biológicas** como DNA, RNA e proteínas.



Exemplo ilustrativo de um alinhamento (GONÇALVES, Macks Wendhell, [s.d.].)

- É um problema **NP-difícil**, **com alta demanda computacional e de memória**.

Motivação

Existem diferentes estratégias para alinhamento de sequências.

Heurísticas:

- Rápida, mas não necessariamente ótima.
  - MAFFT
  - ClustalW

Exatas:

- Custosa, mas pode garantir a solução ótima.
  - PA-Star1
  - PA-Star2

PA-Star: o programa é uma **versão paralela** baseada no **algoritmo exato A-star (A*), usando CPUs para AMS**.

Versão atual (PA-Star2): otimização da divisão de tarefas em máquinas com **processadores assimétricos** (*Asymmetric Multicore Processors* – AMPs).

UnB

Colaboração

Objetivos

- **Comparar PA-Star1 e PA-Star2 em desempenho e uso de memória, identificando qual a melhor versão.**

- **Avaliar a aplicação da melhor versão em um *workflow* científico para o AMS ótimo:**

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

Metodologia Experimental

- Ambiente computacional:
  - **Supercomputador Santos Dumont (SDumont)**
  - Nós com 2× **Intel Xeon Cascade Lake Gold 6252** (48 núcleos, 384 GB RAM). Para testes de alta memória: nós com 768 GB RAM.

- Dados de sequências:
  - **Arquivos multiFASTA de proteínas do banco BAliBASE** (*Benchmark Alignment dataBASE*).
    - **Iniciais**: glg, 1sbp, 1aboA, 1ac5.
    - **Intermediários**: gal4, 1gpb, arp, 1sesA, 2myr, 2cba, 1hvA, 2ack, actin.

Metodologia: Dados de Entrada

Índice	Arquivos (FASTA)	Tamanho do Arquivo	Nº de Seqs	Menor Seq	Maior Seq	Similaridade
1	glg	2.4 K	5	438 aa	486 aa	26.80%
2	1sbp	1.3 K	5	224 aa	263 aa	12.36%
3	1aboA	335	5	49 aa	80 aa	28.75%
4	1ac5	1.8 K	4	421 aa	483 aa	25.10%
5	gal4	1.9 K	5	335 aa	395 aa	15.80%
6	1gpb	4.0 K	5	796 aa	828 aa	42.60%
7	arp	2.1 K	5	380 aa	418 aa	24.16%
8	1sesA	2.2 K	5	417 aa	442 aa	29.87%
9	2myr	1.7 K	4	340 aa	474 aa	14.94%
10	2cba	1.3 K	5	237 aa	259 aa	22.15%
11	1hvA	928	5	136 aa	199 aa	14.07%
12	2ack	2.4 K	5	452 aa	482 aa	18.82%
13	actin	2.0 K	5	379 aa	395 aa	40.25%

Resultados: PA-Star1 x PA-Star2

Índice	Arquivos (FASTA)	Tempo (PA-Star1)	Tempo (PA-Star2)	Consumo de RAM (PA-Star1)	Consumo de RAM (PA-Star2)	Redução de Tempo	Redução de Memória
1	glg	01m:58s	01m:01s	6,25 GB	5,40 GB	48,30%	13,60%
2	1sbp	01m:31s	49s	4,28 GB	3,30 GB	46,20%	22,90%
3	1aboA	01m:03s	49s	5,08 GB	2,94 GB	22,20%	42,13%
4	1ac5	59s	26s	2,99 GB	1,69 MB	55,90%	99,99%
5	gal4	03h:43m:59s	01h:10m:18s	297,77 GB	249,66 GB	68,60%	16,17%
6	1gpb	01h:01m:12s	21m:50s	132,82 GB	110,15 GB	64,30%	17,07%
7	arp	21m:40s	08m:22s	48,78 GB	39,81 GB	61,40%	18,39%
8	1sesA	12m:41s	05m:20s	32,27 GB	25,72 GB	57,90%	20,30%
9	2myr	05m:17s	02m:01s	19,02 GB	13,16 GB	61,80%	30,81%
10	2cba	04m:07s	01m:57s	12,77 GB	8,71 GB	52,60%	31,79%
11	1hvA	03m:35s	01m:44s	10,38 GB	8,53 GB	51,60%	17,82%
12	2ack	02m:14s	01m:09s	7,90 GB	6,60 GB	48,50%	16,46%
13	actin	01m:49s	57s	6,69 GB	3,84 GB	47,70%	42,60%

- A maior redução no tempo de execução foi de 68,6%.
- Todos os arquivos tiveram redução no consumo de memória, 7 dos 13 superior a 20%.

Resultados: PA-Star1 x PA-Star2

Índice	Arquivos (FASTA)	Tempo (PA-Star1)	Tempo (PA-Star2)	Redução de Tempo
1	glg	01m:58s	01m:01s	48,30%
2	1sbp	01m:31s	49s	46,20%
3	1aboA	01m:03s	49s	22,20%
4	1ac5	59s	26s	55,90%
5	gal4	03h:43m:59s	01h:10m:18s	68,60%
6	1gpb	01h:01m:12s	21m:50s	64,30%
7	arp	21m:40s	08m:22s	61,40%
8	1sesA	12m:41s	05m:20s	57,90%
9	2myr	05m:17s	02m:01s	61,80%
10	2cba	04m:07s	01m:57s	52,60%
11	1hvA	03m:35s	01m:44s	51,60%
12	2ack	02m:14s	01m:09s	48,50%
13	actin	01m:49s	57s	47,70%

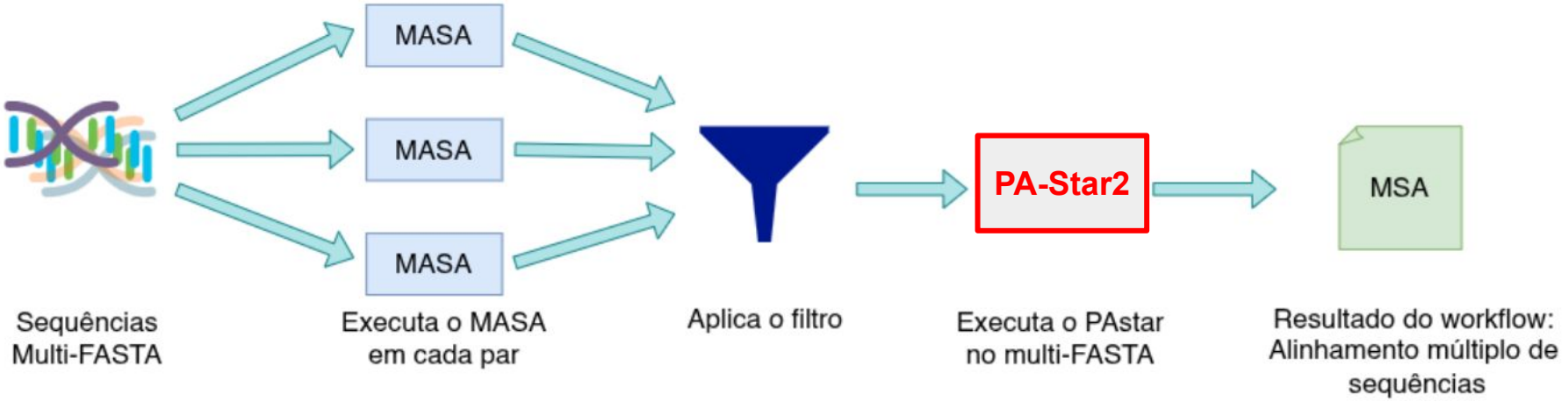
- Na maioria dos casos, redução no tempo de execução de 50% ou mais. A maior redução foi de 68,6%. **Redução média do tempo de 53,6%.**

Resultados: PA-Star1 x PA-Star2

Índice	Arquivos (FASTA)	Consumo de RAM (PA-Star1)	Consumo de RAM (PA-Star2)	Redução de Memória
1	glg	6,25 GB	5,40 GB	13,60%
2	1sbp	4,28 GB	3,30 GB	22,90%
3	1aboA	5,08 GB	2,94 GB	42,13%
4	1ac5	2,99 GB	1,69 MB	99,99%
5	gal4	297,77 GB	249,66 GB	16,17%
6	1gpb	132,82 GB	110,15 GB	17,07%
7	arp	48,78 GB	39,81 GB	18,39%
8	1sesA	32,27 GB	25,72 GB	20,30%
9	2myr	19,02 GB	13,16 GB	30,81%
10	2cba	12,77 GB	8,71 GB	31,79%
11	1hvA	10,38 GB	8,53 GB	17,82%
12	2ack	7,90 GB	6,60 GB	16,46%
13	actin	6,69 GB	3,84 GB	42,60%

- Todos os arquivos tiveram redução no consumo de memória, 7 dos 13 superior a 20%. **Redução média no consumo de RAM de 28,46%.**

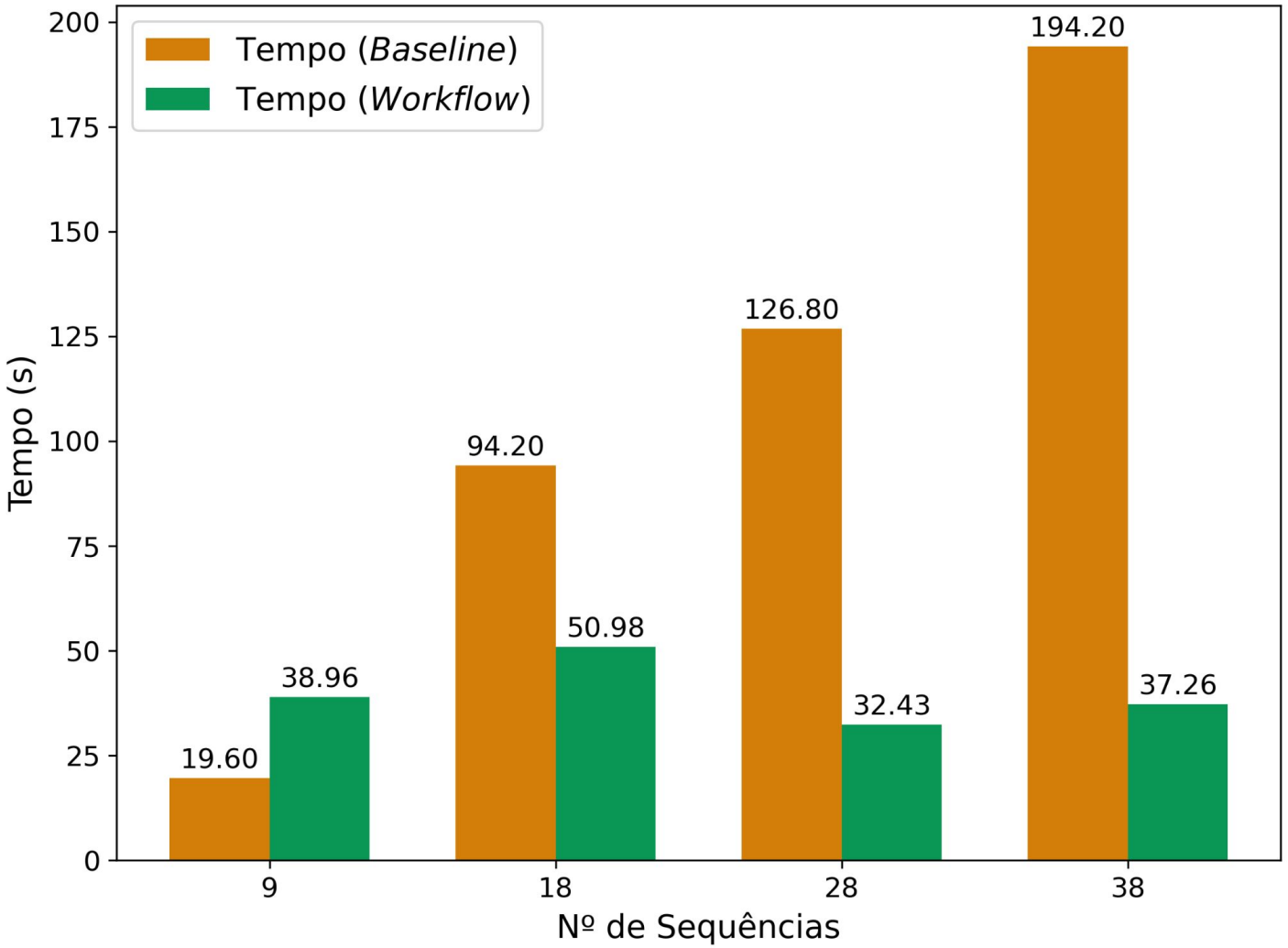
Diagrama básico da aplicação do programa PA-Star2 em um *workflow* científico



Adaptado de Terra, et al. (2025)

Implementado em PyCOMPSs
(*framework* de gerenciamento de tarefas paralelas)

Aplicação em um *workflow* científico



Conclusão

- PA-Star1 x PA-Star2: na maioria dos casos, o PA-Star2 teve **redução em 50% ou mais no tempo de execução dos alinhamentos.**

- **Ainda apresenta limitações no AMS** com grandes conjuntos de sequências **(no máximo 9 sequências)**.

- O **PA-Star2 foi escolhido para integrar um *workflow*,** proposto por **Terra et al. (2025)**, demonstrando ser **viável o AMS mesmo com conjuntos significativos iniciais de sequências.**

Etapas futuras: ampliar as análises com **outras bases de sequências biológicas** e **dados de maior complexidade**, além de comparar o uso do PA-Star2 em um *workflow* científico com **outras aplicações de algoritmos exatos na literatura**.

Referências Bibliográficas

Terra, R., Souza, K., Rocha, H., Osthoff, C., Carvalho, D., Ocaña, K. (2025). Workflow para alinhamento exato de sequências em sistemas de processamento de alto desempenho. *In SSCAD 2025 – Simpósio de Sistemas Computacionais de Alto Desempenho* (trilha principal). [Artigo em apresentação]

Sundfeld, D., Teodoro, G., Melo, 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)*, pp. 146–153, Torino, Italy.

Sundfeld, D., Razzolini, C., Teodoro, G., Boukerche, A., Melo, 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.

Tejedor, E., Becerra, Y., Alomar, G., Queralt, A., et al. (2017). PyCOMPSs: Parallel computational workflows in Python. *International Journal of High Performance Computing Applications*, 31(1), 66–82.

Lima, D. S. (2017). Alinhamento primário e secundário de sequências biológicas em arquiteturas de alto desempenho*. PhD thesis, Universidade de Brasília. Tese (Doutorado em Informática).*

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

Obrigada!

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

[1]Laboratório Nacional de Computação Científica (LNCC) – RJ/Brasil

[2]Faculdade de Educação Tecnológica do Rio de Janeiro (FAETERJ - Petrópolis) – RJ/Brasil

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