

Avaliação de Desempenho Paralelo de Montadores de DNA Velvet e SOAPdenovo2

Evaldo Bezerra da Costa

Programa de Pós Graduação em Informática, PPGI, UFRJ
21941-590, Rio de Janeiro, RJ
E-mail: evaldo.costa@ppgi.ufrj.br

Gabriel P. Silva Marcello Goulart Teixeira

Departamento de Ciência da Computação, DCC, UFRJ
21941-590, Rio de Janeiro, RJ
E-mail: {gabriel, marcellogt}@dcc.ufrj.br

RESUMO

Em bioinformática, existem vários programas disponíveis para montagem de sequência de DNA [1]. Isso geralmente é uma tarefa muito demorada, uma vez que essas sequências de DNA podem ser muito longas e complexas [2]. Assim, uma das principais preocupações com os montadores de DNA é o seu funcionamento e desempenho em sistemas paralelos, de tal maneira que a montagem possa ser realizada em um período de tempo razoável. Neste trabalho, avaliamos o desempenho paralelo dos montadores Velvet e SOAPdenovo2.

Velvet [3] e SOAPdenovo2 [4] são os dois programas de montagem NGS (Sequenciamento de Nova Geração) desenvolvidos usando o método de grafo de De Bruijn e com suporte a computação paralela de multithreaded.

Os testes foram realizados em um servidor com dois processadores Intel Xeon E5-2650 (2,00 GHz, 8 núcleos cada, 20 MB de cache), com 64 GB de memória. Ambos os montadores, Velvet e SOAPdenovo2, foram compilados usando o Intel C++ Compiler Suite XE 13.1 Update 2 para prover um melhor desempenho.

Os dados utilizados para realizar os testes foram fornecidos pelo Instituto de Biologia da UFRJ e foram armazenados em um disco SSD (unidade de estado sólido) local de alta velocidade. O sistema operacional utilizado foi a distribuição Linux CentOS de 64 bits versão 6.3. Os arquivos de dados foram separados em quatro grupos, chamados muriqui8, muriqui9, muriqui10 e muriqui12 conforme apresentado na Tabela I.

Conjunto de dados			
muriqui8	muriqui9	muriqui10	muriqui12
1.7 GB	6.24 GB	16.6 GB	10.8 GB

Tabela I. Tamanho dos conjuntos de dados

Para os resultados apresentados foram realizadas três séries de testes e calculado o tempo médio de execução dos montadores Velvet e SOAPdenovo2. A quantidade de threads usadas foi variada para determinar o speedup de cada montador.

Na Figura 1 é visto que o desempenho Velvet usando conjunto de dados muriqui9 foi melhor. O SOAPdenovo2 obteve um melhor desempenho utilizando o conjunto de dados muriqui10 como apresentado na Figura 2. Os resultados obtidos com o SOAPdenovo2 foram superiores ao Velvet, principalmente em função das melhorias que foram implementadas na versão atual.

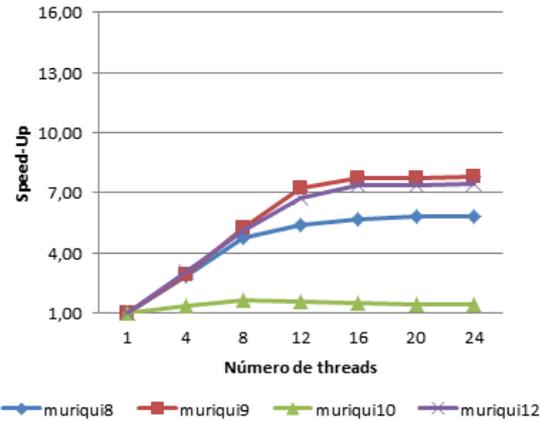


Figura 1: Speed up Velvet

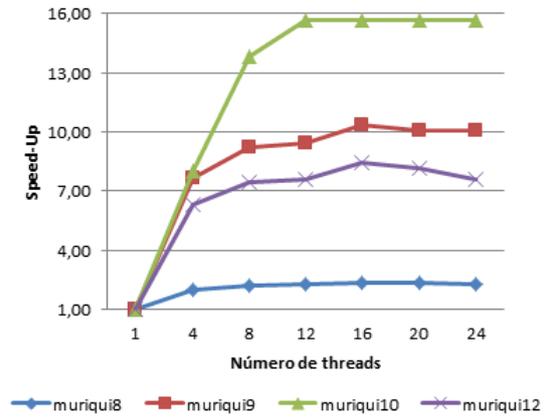


Figura 2: Speed up SOAPdenovo2

O tamanho da sequência a ser montada está relacionada diretamente com o desempenho dos montadores. Embora ambos os montadores utilizassem multithreads em seu processo de execução, o Velvet não utiliza todas as threads especificadas para execução em seu processo, principalmente durante o passo de leitura das sequências. O SOAPdenovo2 utiliza todas as threads especificadas em todos os passos.

Palavras-chave: *Montador DNA, SOAPdenovo2, Velvet, Avaliação de desempenho*

Referências

- [1] Costa, Gustavo Gilson Lacerda. O Laboratório Central de Tecnologias de Alto Desempenho em Ciências da Vida (LaCTAD). Introdução à montagem de genomas, 2013.
- [2] Ruiqiang Li, Hongmei Zhu, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–72, February 2010.
- [3] Zerbino, Daniel and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821-829, 2008.
- [4] Luo R, Liu B, et al. Software and supporting material for “SOAPdenovo2: an empirically improved memory efficient short read de novo assembly”. 2012. 3.