Avaliação de Montadores Paralelos de DNA

Evaldo B. Costa, Gabriel P. Silva

[1]Departamento de Ciência de Computação – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ – Brasil

evaldo.costa@ppgi.ufrj.br, gabriel@dcc.ufrj.br

Resumo. *Esse trabalho faz a avaliação de desempenho de montadores paralelos de sequências de DNA, usados para a montagem de sequências "de novo". Foram avaliados aspectos como ganho (*speedup*), escalabilidade e tempo de execução com diversas entradas. Particularmente, avaliou-se o desempenho com um conjunto de dados conhecido, no caso um cromossoma do genoma humano e também com um conjunto "de novo"de dados usados no mapeamento do genoma completo do primata Muriqui, natural da mata atlântica.*

1. Introdução

Com a descoberta e o aperfeiçoamento de novas técnicas de sequenciamento de DNA, houve um crescimento expressivo na quantidade de dados, o que passou a exigir servidores com maior poder de processamento e capacidade de armazenamento para o processamento desses dados [Costa et al. 2015].

Após o processo de sequenciamento do DNA em equipamentos especializados, é aplicado outro processo chamado de montagem, que é a reconstrução do genoma a partir dos dados gerados. Existem vários programas desenvolvidos para a montagem dessas sequências, os quais possuem versões que executam esse processamento em paralelo.

2. Processo de Sequenciamento e Montagem

Existem dois tipos de processo de montagem: por referência ou "de novo". No processo de montagem por referência, a sequência é comparada com uma já existente, com isso a utilização de recursos computacionais é menor. Quando não se tem um genoma de referência para ser utilizado durante o processo de montagem, é necessário realizar a montagem "de novo". Para se obter bons resultados com o sequenciamento "de novo", é necessário o uso de grande quantidade de processamento e memória [Lander et al. 2001].

A maioria dos novos programas de montagem pode ser classificada como NGS (Next-Generation Sequencing). Com esse método, é possível sequenciar DNA em grande escala e reduzir os custos associados usando o método de sequências curtas [Miller et al. 2010].

3. Resultados

Para os resultados apresentados neste estudo, foram executadas três séries de testes, variando-se a quantidade de *threads* utilizadas.Em seguida, o tempo médio das séries foi calculado para definir-se o tempo de execução e *speedup* obtidos.

Dos montadores avaliados em nosso estudo, os que obtiveram melhor desempenho em termos de *speedup*, independentemente do tamanho dos dados processados, foram Allpaths-LG e SOAPdenovo2 (Figuras 1 e 2).

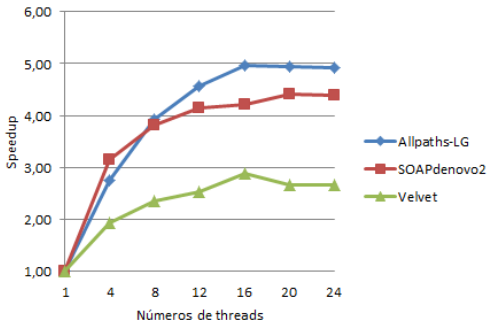

Figura 1. Cromossomo H. 14

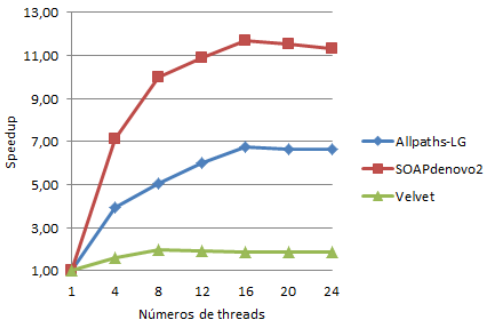

Figura 2. Primata Muriqui

Considerando-se o tempo de execução, o SOAPdenovo2 foi o montador com menor tempo de execução em relação aos demais montadores. Outra análise realizada foi o uso de memória de cada montador. O montador Allpaths-LG utiliza os recursos de memória com maior eficiência. Assim como o montador Allpaths-LG, o SOAPdenovo2 também não necessita utilizar o espaço de *swap* durante o processo de montagem. O montador Velvet utiliza toda a memória disponível durante o processo de montagem, de modo que para continuar o processo de montagem dos dados, precisou fazer uso de área *swap* em disco, com grande impacto negativo no seu desempenho.

4. Conclusão

O desempenho dos montadores Allpaths-LG, SOAPdenovo2 e Velvet foi avaliado com o uso de diversos conjuntos de dados e com o uso de um variado número de *threads*. Os resultados obtidos indicam que esses montadores têm desempenho variado em função do tamanho do conjunto de dados e dos recursos computacionais disponíveis para o processo de montagem. A escalabilidade dos montadores avaliados é menor quando são utilizados conjuntos de dados de menor tamanho, e melhora na medida em que o tamanho desses conjuntos de dados aumenta. Em nossa avaliação, isso ocorre por conta de uma melhor exploração do paralelismo existente, quando a quantidade de dados a serem processados é maior.

Agradecimentos

Os autores agradecem à Microway Incorporated por fornecer os recursos computacionais utilizados para a realização deste trabalho.

Referências

Costa, E. B., Silva, G. P., and Teixeira, M. G. (2015). Performance evaluation of parallel genome assemblers. In Saeed, F. and Haspel, N., editors, *Proc. of the 7th Int. Conf. on Bioinformatics and Computational Biology (BICOB 2015)*, volume 1, pages 31–38. ISCA.

Lander, E. S., Linton, L. M., and Scherer, S. E. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Miller, J., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327.