Uma ferramenta baseada em algoritmos genéticos para a ordenação de montagens parciais de genomas

Vivian M. Y. Pereira[1] **, Camila I. Costa**[1] **, Luciano A. Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
Av. Arlindo Béttio, Ermelino Matarazzo – 03828-000 – São Paulo – SP – Brasil

digiampietri@usp.br

Abstract. *This paper presents a genetic algorithm based approach for sorting two draft genomes. The parameters of the genetic algorithm were calibrated using real data in order to identify the values that optimize the identification of the most appropriate sorting in order to help in subsequent activities of correction or completion of assemblies.*

Resumo. *Este artigo apresenta uma ferramenta baseada em algoritmos genéticos para a ordenação de duas montagens parciais de genomas. Os diferentes parâmetros do algoritmo genético desenvolvido foram calibrados utilizando dados reais de forma a identificar os valores que otimizam a identificação da ordenação mais adequada de forma a auxiliar atividades subsequentes de correção ou complementação das montagens.*

1. Introdução e Motivação

Um dos campos mais conhecidos da bioinformática é o relacionado à microbiologia computacional ou, em particular, à montagem e anotação de genomas. A partir da década de 1970 foram desenvolvidas e aprimoradas técnicas para o sequenciamento de DNA. Até o final da década de 1990, o sequenciamento, montagem e anotação do genoma de uma única bactéria, cujo genoma é tipicamente composto por poucos milhões de pares de bases, era uma tarefa cara e custosa (tanto financeiramente quanto no tempo necessário para ser realizada) [Setubal e Meidanis 1997]. Com os sequenciadores de alto desempenho desenvolvidos nos últimos anos se tornou possível a obtenção de grande volume de DNA (dezenas de milhões de bases) em um único sequenciamento.

O uso de sequenciadores de alta capacidade acarretaram em um grande crescimento no número de genomas sequenciados e, em especial, genomas parcialmente sequenciados (isto é, cuja sequência completa de DNA não é conhecida). Dentre os diversos desafios oriundos deste grande crescimento, este artigo trata do problema de dadas duas montagens parciais de genomas, como ordená-las de forma a facilitar atividades subsequentes de correção ou complementação das montagens. Ao se ter duas corridas de sequenciamento, alguém poderia perguntar se não seria vantajoso juntá-las antes da montagem (obtendo uma única montagem) ao invés de tentar ordenar duas montagens. Porém existem duas situações em que se juntar duas corridas pode não ser interessante ou viável: (i) existem muitos genomas parciais (as vezes chamados de *drafts*) cujas montagens estão disponíveis, porém os *reads* (sequências básicas oriundas dos sequenciadores) não estão disponíveis; (ii) dependendo da tecnologia utilizada para o sequenciamento (especialmente quando cada corrida foi gerada a partir de uma tecnologia diferente) há

diversos relatos que a junção das corridas pode confundir o montador mais do que auxiliar na obtenção da sequência consenso [Mende et al. 2012]. É justamente pensando nestas duas situações que a ferramenta desenvolvida neste artigo foi concebida.

O problema de ordenar duas montagens parciais é computacionalmente complexo, pois dadas duas montagens, uma com n contigs e outra com m contigs existem $n!$ possibilidades de ordenação da primeira montagem e $m!$ possibilidades de ordenação dos contigs da segunda. Já que, para o problema abordado uma ordenação influencia a outra, existirão $n! * m!$ possibilidades de solução, o que é computacionalmente intratável mesmo para valores relativamente pequenos de n e m, assim soluções heurísticas (como o uso de algoritmos genéticos) são necessárias.

O restante deste artigo está organizado da seguinte maneira. A Seção 2 sumariza os trabalhos correlatos. A Seção 3 contém a descrição da metodologia. A Seção 4 apresenta a ferramenta desenvolvida. Por fim, a Seção 5 contém as conclusões.

2. Trabalhos Correlatos

O problema tratado neste artigo de ordenar todas as partes (*contigs*) de duas montagens parciais não é muito tratado na literatura, porém há diversos trabalhos que abordam problemas correlatos. Um problema que já foi amplamente tratado é o uso de um genoma de referência para auxiliar na montagem de um novo genoma [Dias et al. 2012, Digiampietri et al. 2005]. Para este problema, o genoma de referência é, tipicamente, um genoma completo bastando assim reorganizar os *contigs* do genoma que está sendo montado.

Outro trabalho relevante apesar de tratar de um problema diferente, é o trabalho de Mende *et al.* [Mende et al. 2012]. O intuito deste trabalho foi comparar as montagens utilizando dados de diferentes sequenciadores. Foi constatado que, especialmente ao se utilizar sequenciadores que produzem *reads* menores, a capacidade de montagem ficou bastante limitada para projetos de metagenoma. Este tipo de limitação pode ser ampliada ao se combinar *reads* sequenciados utilizando diferentes tecnologias e este foi um dos motivadores deste artigo.

3. Metodologia

A metodologia deste artigo foi composta por três atividades: estudo da literatura correlata, desenvolvimento da ferramenta, calibração dos parâmetros utilizando-se dados reais e validação dos resultados da ferramenta também utilizando dados reais.

4. Ferramenta Desenvolvida e Discussão dos Resultados

A sumarização do problema tratado neste artigo é: dados os contigs de duas montagens parciais do genoma de uma mesma espécie, encontrar uma ordenação para estes contigs de forma a auxiliar atividades subsequentes de correção ou complementação das montagens.

Para isto, a ferramenta desenvolvida é composta de três módulos: Módulo de Alinhamento; Módulo de Ordenação; e Módulo de Visualização.

O **Módulo de Alinhamento** é responsável por identificar os alinhamentos locais entre as sequências dos contigs das duas montagens. Estes alinhamentos serão utilizados como informação básica para o Módulo de Ordenação. Para encontrar os alinhamentos

este módulo inicialmente executa a ferramenta blast[1] e em seguida, dada a lista de alinhamentos em ordem decrescente de *score*, concatena os alinhamentos próximos e exclui os alinhamentos redundantes (que envolvam sequências que já estão presentes em alinhamentos anteriores).

O **Módulo de Ordenação** é o principal módulo do sistema desenvolvido e utiliza um algoritmo genético para ordenar os contigs das duas montagens. A função objetivo do algoritmo visa a ordenar os contigs de forma a minimizar a distância entre os alinhamentos recebidos como entrada. Detalhes dos parâmetros utilizados serão discutidos adiante.

O **Módulo de Visualização** é responsável por apresentar de maneira gráfica o resultado da ordenação indicando também quais são os alinhamentos produzidos pelo Módulo de Alinhamento de forma a facilitar o entendimento da solução encontrada para o usuário. A Figura 1 apresenta a imagem produzida em um estudo de caso real.

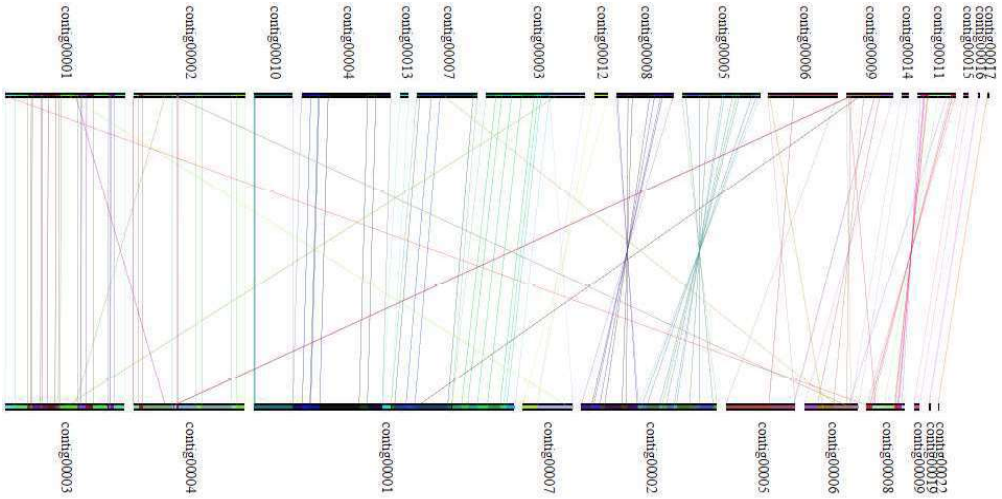
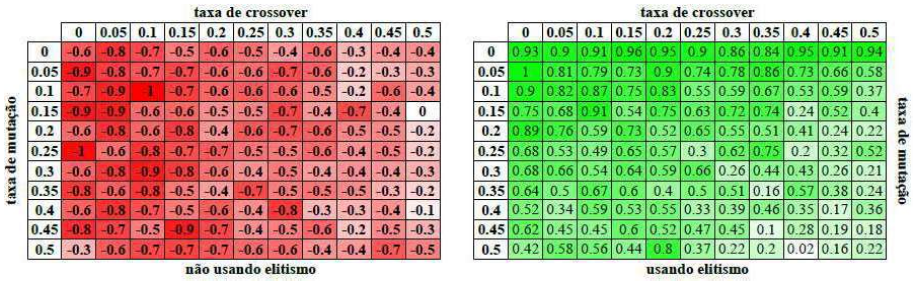


Figura 1. Ordenação de contigs encontrada pela ferramenta desenvolvida

A ferramenta de ordenação possui cinco parâmetros relacionados ao algoritmo genético que podem ser alterados pelo usuário: *tamanho da população inicial*, *número de gerações*, *taxa de crossover*, *taxa de mutação* e *elitismo*. O *número de gerações* pode ser configurado manualmente ou pode-se deixar o algoritmo rodar até que haja estabilidade nos valores da função objetivo. Para os três últimos parâmetros foram executados testes com dados reais (o usuário ainda pode alterar estes valores, porém a configuração original do sistema foi calibrada usando estes dados). O *elitismo* pode estar ou não ativado. Estar ativado significa que os indivíduos melhor ranqueados pela função objetivo terão mais chance de sobreviverem (sofrendo potencialmente mutações) nas gerações futuras. Para as taxas *de mutação* e *de crossover* foram testadas porcentagens de 0 a 50%. Os resultados são apresentados na Figura 2, quanto mais intenso o tom de verde, melhores foram os resultados e quanto mais intenso o tom de vermelho, piores foram os resultados. É fácil observar que os melhores resultados ocorrem usando-se elitismo. Além disso, taxas baixas de mutação e crossover trouxeram os melhores resultados usando-se elitismo. Sem utilizar elitismo, os valores mais altos de crossover produziram melhores resultados (porém inferiores aqueles utilizando-se elitismo). Os resultados foram utilizados para

[1] blast.ncbi.nlm.nih.gov/Blast.cgi

auxiliar no fechamento deste genoma.



taxa de mutação \ taxa de crossover	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0	-0.6	-0.8	-0.7	-0.5	-0.6	-0.5	-0.4	-0.6	-0.3	-0.4	-0.4
0.05	-0.9	-0.8	-0.7	-0.7	-0.6	-0.6	-0.7	-0.6	-0.2	-0.3	-0.3
0.1	-0.7	-0.9	-1	-0.7	-0.6	-0.6	-0.6	-0.5	-0.2	-0.6	-0.4
0.15	-0.9	-0.9	-0.6	-0.6	-0.5	-0.5	-0.7	-0.4	-0.7	-0.4	0
0.2	-0.6	-0.8	-0.6	-0.8	-0.4	-0.6	-0.7	-0.6	-0.5	-0.5	-0.2
0.25	-1	-0.6	-0.8	-0.7	-0.7	-0.5	-0.5	-0.6	-0.4	-0.5	-0.2
0.3	-0.6	-0.8	-0.9	-0.8	-0.6	-0.4	-0.5	-0.4	-0.4	-0.4	-0.3
0.35	-0.8	-0.6	-0.8	-0.5	-0.4	-0.7	-0.5	-0.5	-0.5	-0.3	-0.2
0.4	-0.6	-0.8	-0.7	-0.5	-0.6	-0.4	-0.8	-0.3	-0.3	-0.4	-0.1
0.45	-0.8	-0.7	-0.5	-0.9	-0.7	-0.4	-0.5	-0.6	-0.2	-0.5	-0.3
0.5	-0.3	-0.6	-0.7	-0.7	-0.7	-0.6	-0.6	-0.4	-0.4	-0.7	-0.5

taxa de mutação \ taxa de crossover	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0	0.93	0.9	0.91	0.96	0.95	0.9	0.86	0.84	0.95	0.91	0.94
0.05	1	0.81	0.79	0.73	0.9	0.74	0.78	0.86	0.73	0.66	0.58
0.1	0.9	0.82	0.87	0.75	0.83	0.55	0.59	0.67	0.53	0.59	0.37
0.15	0.75	0.68	0.91	0.54	0.75	0.63	0.72	0.74	0.24	0.52	0.4
0.2	0.89	0.76	0.59	0.73	0.52	0.65	0.55	0.51	0.41	0.24	0.22
0.25	0.68	0.53	0.49	0.65	0.57	0.3	0.62	0.75	0.2	0.32	0.52
0.3	0.68	0.66	0.54	0.64	0.59	0.66	0.26	0.44	0.43	0.26	0.21
0.35	0.64	0.5	0.67	0.6	0.4	0.5	0.51	0.16	0.57	0.38	0.24
0.4	0.52	0.34	0.59	0.53	0.55	0.33	0.39	0.46	0.35	0.17	0.36
0.45	0.62	0.45	0.45	0.6	0.52	0.47	0.45	0.1	0.28	0.19	0.18
0.5	0.42	0.58	0.56	0.44	0.8	0.37	0.22	0.2	0.02	0.16	0.22

Figura 2. Resultado da calibração dos parâmetros

Um estudo de caso foi realizado utilizando-se dois sequenciamentos de *Mycobacterium* (gênero de bactérias cujo genoma costuma possuir em torno de 5 milhões de pares de bases). Cada sequenciamento produziu cerca de 20 contigs com mais de 500 bases em cada. Para este problema, a ordenação de contigs encontrou a solução ótima em 10 segundos[2] (incluindo o tempo de alinhamento). A ordenação resultante pode ser vista na Figura 1.

5. Conclusões

Este artigo apresentou uma ferramenta para a ordenação de montagens parciais de genoma de forma a facilitar atividades de complementação ou fechamento do genoma. A ferramenta desenvolvida foi calibrada e testada com dados reais e está sendo utilizada para auxiliar no fechamento de genomas de bactérias.

Como trabalhos futuros pretende-se desenvolver um módulo para destacar as informações mais relevantes obtidas com a ordenação (por exemplo, as regiões de cada montagem que são complementares a outra montagem).

Agradecimentos

O trabalho apresentado neste artigo foi parcialmente financiado pela FAPESP, pelo CNPq e pelo Programa de Educação Tutorial do MEC.

Referências

Dias, Z., Dias, U., e Setubal, J. (2012). Sis: a program to generate draft genome sequence scaffolds for prokaryotes. *BMC Bioinformatics*, 13(1):96.

Digiampietri, L. A., Perdigueiro, J. M., de Almeida Junior, A. J., Faria, D. M., Ostroski, E. H., Costa, G. G. L., e Perez, M. C. (2005). Fact and task oriented system for genome assembly and annotation. In *Proceedings of the 2005 Brazilian Conference on Advances in Bioinformatics and Computational Biology*, BSB'05, pages 238–241, Berlin, Heidelberg. Springer-Verlag.

Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., Raes, J., e Bork, P. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE*, 7(2):e31386.

Setubal, J. e Meidanis, J. (1997). *Introduction to Computational Molecular Biology*. PWS Publishing Company, Boston.

[2]Os testes foram executados em um notebook com processador Intel Core 2 Duo.