Ordenação de Permutações com Sinais por Reversões e Transposições

Klairton de Lima Brito[1] **e Zanoni Dias (Orientador)**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (Unicamp)
Av. Albert Einstein, 1251 – 13.083-852 – Campinas – SP – Brasil

{klairton,zanoni}@ic.unicamp.br

Abstract. *Finding a genome rearrangements sequence capable of transforming one genome into another can be very useful in comparative genomics. Depending on the scenario in which we come across, the characteristics sought for this genome rearrangements sequence may be different. In this dissertation, we work with genomes in which the orientation of genes is known and we considered the reversal and transposition rearrangement events. We address the classical problem in which both events affect the genome with the same frequency. In addition, we investigated a version of the problem in which the events occur with a different frequency.*

Resumo. *Determinar uma sequência de rearranjos de genomas capaz de transformar um genoma em outro pode ser bastante útil na genômica comparativa. Dependendo do cenário em que nos deparamos as características buscadas para essa sequência de rearranjos de genomas podem ser diferentes. Nessa dissertação, trabalhamos com genomas em que a orientação dos genes é conhecida e consideramos os eventos de rearranjo de genomas de reversão e transposição. Abordamos o problema clássico no qual ambos os eventos afetam o genoma com a mesma frequência. Além disso, investigamos uma variação do problema na qual cada tipo de evento ocorre com uma frequência diferente.*

1. Introdução

Um dos desafios encontrado na genômica comparativa é compreender melhor o processo evolutivo das espécies. Uma forma utilizada para inferir a distância evolutiva entre dois organismos é através da distância de rearranjo de genomas. Rearranjos de genomas são eventos mutacionais que podem afetar grandes porções do DNA [Fertin et al. 2009], sendo a reversão e a transposição os eventos de rearranjo mais estudados na literatura [Bergeron 2005, Bafna and Pevzner 1998]. Uma reversão afeta um segmento do genoma invertendo a posição e a orientação dos genes desse segmento, enquanto uma transposição troca dois segmentos consecutivos do genoma de posição, mas sem afetar a posição e a orientação dos genes dentro dos segmentos. A representação de um genoma pode ser feita de diversas formas. Em particular, quando os genes não apresentam repetições podemos associar para cada gene um número inteiro e a representação do genoma é dada por uma permutação. Quando a orientação dos genes é conhecida cada elemento da permutação recebe um sinal de positivo ou negativo, caso contrário o sinal é omitido. A permutação resultante da representação do genoma é chamada de com sinais e sem sinais quando a orientação dos genes é conhecida e desconhecida, respectivamente.

Um modelo de rearranjo determina o conjunto de eventos de rearranjo permitidos para transformar um genoma em outro. Quando adotamos a representação do genoma na forma de uma permutação podemos simplificar o problema como sendo um problema de ordenação. Nesse caso, queremos alcançar um genoma específico representado por uma permutação em que os elementos estão ordenados de forma crescente. Essa permutação é chamada de identidade, representada por $\iota = (+1 \ \ldots +n)$.

Quando nos deparamos com um cenário em que os eventos de rearranjo ocorrem com uma mesma frequência o objetivo consiste em encontrar a menor sequência de eventos de rearranjo capaz de ordenar uma permutação π. Em outras palavras, transformar π em ι. Considerando um modelo composto apenas pelo evento de reversão e permutações com sinais temos um algoritmo exato em tempo polinomial para realizar essa tarefa [Hannenhalli and Pevzner 1999]. Existe também um algoritmo que executa em tempo linear [Bader et al. 2001] se estivermos interessados somente no tamanho da sequência de reversões. Quando consideramos o mesmo modelo, mas utilizando permutações sem sinais, foi provado que o problema pertence à classe de problemas NP-Difícil [Caprara 1999] e o melhor algoritmo conhecido possui um fator de aproximação 1.375 [Berman et al. 2002]. De maneira similar, quando consideramos apenas o evento de transposição e permutações sem sinais, o problema também pertence à classe de problemas NP-Difícil [Bulteau et al. 2012] e o melhor algoritmo conhecido possui um fator de aproximação 1.375 [Elias and Hartman 2006].

Considerando permutações com e sem sinais em um modelo de rearranjo que permite o uso dos eventos de reversão e transposição, temos que ambos os problemas apresentam uma complexidade desconhecida e os melhores algoritmos conhecidos apresentam um fator de aproximação 2 [Walter et al. 1998] e $2.8334+\epsilon$ [Rahman et al. 2008], respectivamente.

Podemos também nos deparar com outro cenário em que determinados eventos de rearranjo são mais propícios de ocorrer em relação a outros [Blanchette et al. 1996, Yancopoulos et al. 2005]. Nesse caso, uma forma de representar esse comportamento é associar pesos para cada tipo de evento de rearranjo e o objetivo consiste em encontrar a sequência de eventos de rearranjo capaz de ordenar uma permutação π cuja soma total dos pesos seja miníma. Uma transreversão troca dois segmentos consecutivos do genoma de posição sendo que um dos segmentos é invertido, alterando assim a posição e a orientação dos elementos desse segmento. Eriksen [Eriksen 2002] apresentou um algoritmo com fator de aproximação 7/6 considerando um modelo composto pelos eventos de reversão, transposição e transreversão adotando os pesos 1, 2 e 2, respectivamente. Bader e Ohlebusch [Bader and Ohlebusch 2007] apresentaram um algoritmo com fator de aproximação 3/2 para o mesmo problema, mas adotando os pesos 1 para reversões e qualquer valor em um intervalo de 1 até 2 para transposições e transreversões.

As seções a seguir apresentam as principais contribuições da dissertação. A Seção 2 apresenta o trabalho que foi desenvolvido com foco no problema de Ordenação de Permutações com Sinais por Reversões e Transposições. A Seção 3 sumariza os resultados obtidos no âmbito do problema de Ordenação de Permutações com Sinais por Reversões e Transposições Ponderadas. A Seção 4 apresenta as contribuições do trabalho e a Seção 5 conclui este resumo.

2. Ordenação de Permutações com Sinais por Reversões e Transposições

Para esse problema assumimos que os eventos de reversão e transposição ocorrem com a mesma frequência e que a orientação dos genes é conhecida. Nesse caso, queremos encontrar a menor sequência de eventos de rearranjo capaz de ordenar uma permutação π qualquer. O uso de heurísticas na área de Rearranjo de Genomas para melhorar resultados já conhecidos é uma prática frequente [Dias et al. 2014, Dias 2012]. Dessa forma, desenvolvemos três heurísticas para o problema de Ordenação de Permutações com Sinais por Reversões e Transposições. As heurísticas chamam-se *Sliding Window*, *Look Ahead* e *Iterative Sliding Window*. Todas foram desenvolvidas de maneira a possibilitar o uso em qualquer algoritmo criado especificamente para o problema em questão ou que forneça uma solução viável para o mesmo. A partir de uma sequência de ordenação inicial, as heurísticas aplicam técnicas que buscam reduzir a quantidade de eventos utilizados por essa sequência de ordenação. Para verificarmos a eficiência das nossas heurísticas criamos uma base de dados contendo permutações distintas com tamanhos variados e diferentes características. Aplicamos as heurísticas em vários algoritmos conhecidos na literatura e comparamos os resultados obtidos com aqueles fornecidos pelos algoritmos sem nenhuma heurística aplicada. Abordamos três versões do problema, sendo elas:

- Clássica: Os eventos de reversão e transposição não possuem restrição referente a região do genoma que devem afetar.
- Prefixo: Os eventos de reversão e transposição devem afetar a região do início do genoma.
- Prefixo ou sufixo: Os eventos de reversão e transposição devem afetar a região do início ou fim do genoma.

Em comparação com as soluções fornecidas pelos algoritmos sem nenhuma heurística aplicada, na maioria dos casos, foi possível perceber que as heurísticas foram capazes de fornecer soluções de melhor qualidade. Por fim, mostramos como combinar as heurísticas *Look Ahead* e *Sliding Window* visando a obtenção de resultados ainda melhores.

3. Ordenação de Permutações com Sinais por Reversões e Transposições Ponderadas

Para esse problema abordamos um cenário em que a orientação dos genes é conhecida e os eventos de reversão e transposição ocorrem com uma frequência diferente. Uma forma de representar esse comportamento é adotando pesos para cada evento. Nesse caso, queremos encontrar a sequência de eventos de rearranjo capaz de ordenar uma permutação π qualquer e que a soma dos pesos dos eventos utilizados seja mínima. Com base nos estudos realizados por Eriksen [Eriksen 2001] e Bader *et al.* [Bader et al. 2008] utilizamos os pesos 2 e 3 para os eventos de reversão e transposição, respectivamente.

Desenvolvemos quatro algoritmos para o problema com fatores de aproximação 3, 2, $5/3$ e $3/2$. Além disso, com base no algoritmo de aproximação $3/2$, criamos um esquema que garante uma aproximação considerando diferentes cenários de ponderações para os eventos de reversão e transposição. Realizamos experimentos para verificarmos os resultados práticos fornecidos pelos algoritmos. Para isso, criamos uma base de dados contendo permutações com características que representam diferentes cenários evolutivos. Com base nos resultados, observamos que mesmo os algoritmos com fatores de aproximação 3 e 2 foram capazes de fornecer soluções equiparáveis com os algoritmos com fatores de aproximação $5/3$ e $3/2$, na média.

4. Contribuições

Uma versão preliminar apresentando as heurísticas *Sliding Window* e *Look Ahead* e abordando a versão clássica do problema de Ordenação de Permutações com Sinais por Reversões e Transposições foi apresentada na 5th International Conference on Algorithms for Computational Biology (AlCoB 2018) [Brito et al. 2018]. Uma versão estendida contendo todas as heurísticas e abordando as versões clássica, prefixo e prefixo ou sufixo do problema foi submetida para IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) e encontra-se em processo de revisão.

Apresentamos os resultados preliminares obtidos para o problema de Ordenação de Permutações com Sinais por Reversões e Transposições Ponderadas durante o 11th Brazilian Symposium on Bioinformatics (BSB 2018) [Oliveira et al. 2018]. Nesse artigo, mostramos os algoritmos de aproximação com fatores 2 e $5/3$ e os resultados experimentais. Posteriormente, tivemos um artigo aceito para publicação no Journal of Computational Biology (JCB) [Oliveira et al. 2019] contendo uma análise simplificada para o algoritmo com fator de aproximação $5/3$, o algoritmo com fator de aproximação $3/2$ e o esquema de aproximação considerando diferentes cenários de ponderações. A Tabela 1 resume as contribuições resultantes do trabalho. A coluna da esquerda mostra o nome da conferência ou periódico que realizamos a submissão de artigos completos, a coluna do meio apresenta o ano em que o trabalho foi publicado e a última coluna apresenta o Qualis da conferência ou periódico na área de Ciência da Computação.

Tabela 1. Sumarização das produções bibliográficas.

Conferências / Periódicos	**Ano**	**Qualis**
International Conference on Algorithms for Computational Biology	2018	-
Brazilian Symposium on Bioinformatics	2018	B3
Journal of Computational Biology	2019	A1
IEEE/ACM Transactions on Computational Biology and Bioinformatics (em revisão)	-	A2

5. Conclusão

Nesse resumo, apresentamos a dissertação de mestrado defendida por Klairton de Lima Brito [de Lima Brito 2018], que aborda duas variações do problema de Ordenação de Permutações com Sinais por Reversões e Transposições. Em um primeiro cenário, assumimos que os eventos de reversão e transposição ocorrem com a mesma frequência e focamos nossos esforços no desenvolvimentos de heurísticas que pudessem ser aplicas em algoritmos de aproximação visando o refinamento das soluções fornecidas. Desenvolvemos um total de três heurísticas e aplicamos em três versões do problema. Os experimentos mostraram que fomos capazes que melhorar significativamente os resultados fornecidos pelos algoritmos de aproximação conhecidos na literatura nas três versões abordadas. No segundo cenário, abordamos o problema em que os eventos de reversão e transposição ocorrem com uma frequência diferente. Para reproduzirmos esse comportamento adotamos os pesos 2 e 3 para os eventos de reversão e transposição, respectivamente, e buscamos encontrar sequências de ordenação em que a soma dos pesos fosse mínima. Desenvolvemos quatro algoritmos de aproximação e realizamos experimentos para constatarmos os resultados práticos fornecidos de cada um deles.

Referências

Bader, D. A., Moret, B. M. E., and Yan, M. (2001). A Linear-Time Algorithm for Computing Inversion Distance Between Signed Permutations with an Experimental Study. *Journal of Computational Biology*, 8:483–491.

Bader, M., Abouelhoda, M. I., and Ohlebusch, E. (2008). A Fast Algorithm for the Multiple Genome Rearrangement Problem with Weighted Reversals and Transpositions. *BMC Bioinformatics*, 9(1):1–13.

Bader, M. and Ohlebusch, E. (2007). Sorting by Weighted Reversals, Transpositions, and Inverted Transpositions. *Journal of Computational Biology*, 14(5):615–636.

Bafna, V. and Pevzner, P. A. (1998). Sorting by Transpositions. *SIAM Journal on Discrete Mathematics*, 11(2):224–240.

Bergeron, A. (2005). A Very Elementary Presentation of the Hannenhalli-Pevzner Theory. *Discrete Applied Mathematics*, 146(2):134–145.

Berman, P., Hannenhalli, S., and Karpinski, M. (2002). 1.375-Approximation Algorithm for Sorting by Reversals. In *Proceedings of the 10th Annual European Symposium on Algorithms (ESA'2002)*, volume 2461 of *Lecture Notes in Computer Science*, pages 200–210. Berlin/Heidelberg, Germany.

Blanchette, M., Kunisawa, T., and Sankoff, D. (1996). Parametric Genome Rearrangement. *Gene*, 172(1):GC11–GC17.

Brito, K. L., Oliveira, A. R., Dias, U., and Dias, Z. (2018). Heuristics for the Sorting Signed Permutations by Reversals and Transpositions Problem. In *Algorithms for Computational Biology*, volume 10849, pages 65–75. Heidelberg, Germany.

Bulteau, L., Fertin, G., and Rusu, I. (2012). Sorting by Transpositions is Difficult. *SIAM Journal on Computing*, 26(3):1148–1180.

Caprara, A. (1999). Sorting Permutations by Reversals and Eulerian Cycle Decompositions. *SIAM Journal on Discrete Mathematics*, 12(1):91–110.

de Lima Brito, K. (2018). Sorting Signed Permutations by Reversals and Transpositions. Master's thesis, Institute of Computing, University of Campinas, Brazil. In Portuguese.

Dias, U., Galvão, G. R., Lintzmayer, C. N., and Dias, Z. (2014). A General Heuristic for Genome Rearrangement Problems. *Journal of Bioinformatics and Computational Biology*, 12(3):26.

Dias, U. M. (2012). *Problemas de Comparação de Genomas*. PhD thesis, Institute of Computing, University of Campinas. In Portuguese.

Elias, I. and Hartman, T. (2006). A 1.375-Approximation Algorithm for Sorting by Transpositions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4):369–379.

Eriksen, N. (2001). *Combinatorics of Genome Rearrangements and Phylogeny*. Teknologie licentiat thesis, Kungliga Tekniska Högskolan, Stockholm.

Eriksen, N. (2002). $(1+\epsilon)$-Approximation of Sorting by Reversals and Transpositions. *Theoretical Computer Science*, 289(1):517–529.

Fertin, G., Labarre, A., Rusu, I., Tannier, É., and Vialette, S. (2009). *Combinatorics of Genome Rearrangements*. Computational Molecular Biology. The MIT Press, London, England.

Hannenhalli, S. and Pevzner, P. A. (1999). Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals. *Journal of the ACM*, 46(1):1–27.

Oliveira, A. R., Brito, K. L., Dias, Z., and Dias, U. (2018). Sorting by Weighted Reversals and Transpositions. In *Advances in Bioinformatics and Computational Biology*, pages 38–49. Heidelberg, Germany.

Oliveira, A. R., Brito, K. L., Dias, Z., and Dias, U. (2019). Sorting by Weighted Reversals and Transpositions. *Journal of Computational Biology*, pages 1–12.

Rahman, A., Shatabda, S., and Hasan, M. (2008). An Approximation Algorithm for Sorting by Reversals and Transpositions. *Journal of Discrete Algorithms*, 6(3):449–457.

Walter, M. E. M. T., Dias, Z., and Meidanis, J. (1998). Reversal and Transposition Distance of Linear Chromosomes. In *Proceedings of the 5th International Symposium on String Processing and Information Retrieval (SPIRE'1998)*, pages 96–102, Los Alamitos, CA, USA.

Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient Sorting of Genomic Permutations by Translocation, Inversion and Block Interchange. *Bioinformatics*, 21(16):3340–3346.