Uma Ferramenta de Auditoria para Algoritmos de Rearranjo de Genomas

Gustavo Rodrigues Galvão (Autor)[1], **Zanoni Dias (Orientador)**[1]

[1]Instituto de Computação
Universidade Estadual de Campinas (Unicamp) – Campinas, SP – Brasil

{ggalvao,zanoni}@ic.unicamp.br

Abstract. *This paper presents the results from the investigation of genome rearrangement algorithms, which find application in the estimation of evolutionary distance between species. The great majority of these algorithms are approximate, therefore we developed a tool for evaluating their results. We implemented and evaluated 16 approximate genome rearrangement algorithms using this tool. As a result, we showed that the approximation ratios of some of these algorithms are tight, contradicting some hypotheses raised in the literature, and we conjectured that the approximation ratios of some others also are.*

Resumo. *Este artigo apresenta os resultados da investigação de algoritmos de rearranjo de genomas, que possuem aplicações na estimação da distância evolutiva entre espécies. A grande maioria desses algoritmos são aproximados, por isso nós desenvolvemos uma ferramenta para avaliar suas respostas. Implementamos e avaliamos 16 algoritmos de rearranjo de genomas aproximados utilizando esta ferramenta. Como resultado, mostramos que os fatores de aproximação de alguns desses algoritmos são justos, contradizendo algumas hipóteses levantadas na literatura, e conjecturamos que o fator de aproximação de alguns outros também são.*

1. Introdução

Ao longo da evolução, mutações globais podem alterar a ordem dos genes de um genoma. Tais mutações são chamadas de eventos de rearranjo. Em Rearranjo de Genomas, estimamos a distância evolutiva entre dois genomas calculando-se a distância de rearranjo entre eles, que é o tamanho da menor sequência de eventos de rearranjo que transforma um genoma no outro. Essa distância, por sua vez, pode ser utilizada para estimar a distância evolutiva entre espécies.

Representando genomas como permutações, nas quais os genes aparecem como elementos, a distância de rearranjo pode ser obtida resolvendo-se o problema combinatório de ordenar uma permutação utilizando o menor número de eventos de rearranjo. Este problema, que é referido como Problema da Ordenação por Rearranjo, varia de acordo com o modelo de rearranjo sendo considerado, que nada mais é do que os tipos de eventos de rearranjo permitidos para ordenar uma permutação.

Em nosso trabalho, focamos nosso estudo em dois tipos de eventos de rearranjo: reversões e transposições. Variações do Problema da Ordenação por Rearranjo que consideram esses eventos têm se mostrado difíceis de serem resolvidas otimamente, por isso

a maior parte dos algoritmos propostos – os quais denominamos genericamente por algoritmos de rearranjo de genomas – são aproximados e é esperado que os próximos avanços ocorram nesse sentido. Esse fato nos motivou a construir uma ferramenta para auxiliar a avaliação das respostas desse tipo de algoritmo.

2. Objetivos

Um método empregado na literatura para avaliar as respostas de um algoritmo de rearranjo de genomas que não produz respostas garantidamente ótimas é o que chamamos de *auditoria*. Ele consiste em calcular a distância de rearranjo de todas as permutações que possuem até um certo tamanho, compará-las com as respostas retornadas pelo algoritmo para essas permutações e produzir estatísticas que meçam a qualidade das respostas.

O processo de auditoria consome uma quantidade considerável de tempo e esforço, em sua maioria gastos implementando um algoritmo que calcula as distâncias de rearranjo e, em seguida, computando as distâncias de rearranjo. Sendo assim, uma ferramenta de auditoria poderia mitigar o tempo e o esforço gastos para realizar ambas as tarefas, tornando assim o processo de auditoria mais rápido e menos laborioso. Além de facilitar esse processo, tal ferramenta poderia promover uma certa padronização da avaliação de algoritmos de rearranjo de genomas. A razão disso é que, se todos algoritmos fossem avaliados utilizando um mesmo método (ou um conjunto de métodos), tornar-se-ia muito mais fácil compará-los, sendo inclusive suficiente avaliá-los apenas uma vez.

Um aspecto teórico não muito explorado nos trabalhos que apresentam algoritmos de rearranjo de genomas aproximados é a justeza do fator de aproximação. Para afirmar que o fator de aproximação de um algoritmo aproximado é justo, faz-se necessário demonstrar que existe uma infinidade de instâncias de entrada para as quais a razão entre a resposta do algoritmo e a solução ótima é igual ao fator de aproximação do algoritmo. Além de servirem para mostrar que não é possível melhorar o fator de aproximação de um algoritmo aproximado, as instâncias de entrada que demonstram a justeza do fator de aproximação oferecem uma visão crítica de como o algoritmo funciona e frequentemente levam à ideias que dão origem a algoritmos com fatores de aproximações melhores.

Deste modo, o trabalho foi realizado tendo-se em vista dois objetivos principais:

- construir uma ferramenta de auditoria para algoritmos de rearranjo de genomas;
- avaliar algoritmos de rearranjo de genomas aproximados com a ferramenta e discutir a justeza do fator de aproximação desses algoritmos com base nos resultados obtidos, trazendo para o centro da discussão um tópico pouco abordado na literatura de Rearranjo de Genomas.

3. Resultados e Contribuições

Para construir a ferramenta de auditoria, a qual chamamos de GRAAu [Galvão and Dias 2012a], calculamos a distância de rearranjo de todas as permutações sem sinal com até 13 elements e de todas as permutações com sinal com até 10 elementos com respeito a diversos modelos de rearranjo abordados na literatura que consideram reversões ou transposições. No melhor do nosso conhecimento, essa foi a primeira vez que um cálculo desse tipo foi realizado para permutações contendo esses números de elementos. Esse feito se deve, em grande parte, ao fato de termos desenvolvido um algoritmo de busca em largura simples e flexível que é mais eficiente em termos

de uso de memória do que qualquer outro algoritmo que encontramos na literatura [Galvão and Dias 2011a, Galvão and Dias 2011b]. Além disso, para melhorar o tempo de execução, que é exponencial no tamanho das permutações, criamos uma maneira de paralelizá-lo.

Analisando a distribuição das distâncias de rearranjo, pudemos observar alguns fatos interessantes. Em um primeiro momento, procuramos por outros trabalhos que também apresentassem as distribuições das distâncias de rearranjo para que pudéssemos confrontar com aquelas calculadas por nós. Como resultado, descobrimos que a distribuição da distância de reversão apresentada por Kececioglu e Sankoff [Kececioglu and Sankoff 1995] não está correta [Galvão and Dias 2011a]. Em um segundo momento, procuramos por conjecturas a respeito do diâmetro, que é o maior número de eventos de rearranjo necessários para ordenar uma permutação considerando-se todas as permutações com o mesmo número de elementos. Como resultado, verificamos que a conjectura de Dias e Meidanis [Dias and Meidanis 2002] a respeito do diâmetro de transposição de prefixo é válida e que a conjectura de Walter, Dias e Meidanis [Walter et al. 1998] a respeito do diâmetro de reversão com sinal e transposição é inválida. Em razão disso, apresentamos uma nova conjectura [Galvão and Dias 2011c].

Em uma tentativa de melhor capturar a maneira como as distâncias de rearranjo estão distribuídas, propusemos duas medidas, o diâmetro transversal e a longevidade, e apresentamos algumas conjecturas a respeito delas [Galvão and Dias 2011c].

A fim de que pudéssemos ilustrar as aplicações do GRAAu, implementamos e auditamos 16 algoritmos de rearranjo de genomas aproximados relativos a 6 variações do Problema da Ordenação por Rearranjo: Problema da Ordenação por Reversões, Problema da Ordenação por Reversões de Prefixo, Problema da Ordenação por Reversões de Prefixo com Sinal, Problema da Ordenação por Reversões Curtas, Problema da Ordenação por Transposições e o Problema da Ordenação por Transposições de Prefixo. Dos 16 algoritmos, 4 deles foram propostos por nós: dois para o Problema da Ordenação por Reversões de Prefixo com Sinal, um para o Problema da Ordenação por Transposições e um para o Problema da Ordenação por Transposições de Prefixo.

A implementação dos algoritmos dependeu de uma revisão aprofundada dos conceitos relativos a cada um deles. Tal revisão acabou resultando em algumas contribuições, dentre as quais destacamos:

- a complementação da demonstração do fator de aproximação de um algoritmo 3-aproximado para o Problema da Ordenação por Transposições [Galvão and Dias 2012c];
- e a demonstração de um fator de aproximação igual a 3 para uma versão restrita de uma heurística para o Problema da Ordenação por Transposições [Galvão and Dias 2012c].

Uma aplicação para as estatísticas produzidas pelo GRAAu é a comparação. Utilizando as estatísticas obtidas para os algoritmos de rearranjo de genomas auditados, conseguimos, na maioria dos casos, obter resultados claros quanto à superioridade de um algoritmo em relação aos outros.

Outra aplicação das estatísticas produzidas pelo GRAAu é a validação. Quanto a isso, discutimos principalmente a questão da justeza do fator de aproximação, que é um

aspecto praticamente não explorado pela literatura de Rearranjo de Genomas. Baseado nas informações produzidas pelo GRAAu, demonstramos que o fator de aproximação de 7 dos 16 algoritmos aproximados considerados é justo [Galvão and Dias 2012d, Galvão and Dias 2012b]. Ademais, conjecturamos que o fator de aproximação de outros 6 algoritmos também é justo e levantamos a hipótese de que o fator de aproximação de um outro algoritmo pode ser diminuído. Assim, dos 16 algoritmos considerados, conseguimos demonstrar ou inferir algum resultado quanto à justeza do fator de aproximação para 14 algoritmos, o que evidencia a eficácia da ferramenta nesse sentido.

Os resultados referentes à justeza do fator de aproximação contrapõem duas hipóteses encontradas na literatura. Uma das hipóteses, levantada por Benoît-Gagné e Hamel [Benoît-Gagné and Hamel 2007], é de que o fator de aproximação do algoritmo 3-aproximado desenvolvido por eles para o Problema da Ordenação por Transposições "tende a um número significantemente menor do que 3". Nossos resultados apontaram para uma direção contrária, indicando que o fator de aproximação desenvolvidos por eles é justo [Galvão and Dias 2011d, Galvão and Dias 2012c].

Outra hipótese, levantada por Fischer e Ginzinger [Fischer and Ginzinger 2005], é a de que o fator de aproximação do algoritmo 2-aproximado desenvolvido por eles para o Problema da Ordenação por Reversões de Prefixo pode ser diminuído. Apesar deles não terem especificado um algoritmo, nós implementamos dois algoritmos que utilizam a estratégia gulosa proposta por eles e mostramos que o fator de aproximação de um deles é justo e que o fator de aproximação do outro dá claras evidências de que também é justo. Isso significa que, no pior caso, tal estratégia não é capaz de produzir um algoritmo aproximado com um fator de aproximação menor do que 2 [Galvão and Dias 2012d].

Finalmente, mostramos que os resultados experimentais apresentados por Walter, Dias e Meidanis [Walter et al. 2000] a respeito de um algoritmo 2.25-aproximado para o Problema da Ordenação por Transposições não são corretos [Galvão and Dias 2012c]. Apesar dessa conclusão não ter sido derivada das estatísticas produzidas pelo GRAAu, ela poderia ter sido, isto é, se tivéssemos auditado a implementação deles com o GRAAu, certamente obteríamos resultados que não estariam de acordo com o limite superior derivado na teoria e, consequentemente, detectaríamos o erro.

3.1. Publicações

As contribuições apresentadas na dissertação foram publicadas, quase na sua totalidade, em anais de conferências nacionais e internacionais de biologia computacional, que são detalhadas a seguir:

- *ACM Symposium on Applied Computing - Conference Track on Bioinformatics and Computational Systems Biology*: Conferência patrocinada pelo SIGAPP (*Special Interest Group on Applied Computing*) da ACM. Ela foi classificada pela CAPES como Qualis A1 em 2012.
- *Brazilian Symposium on Bioinformatics*: Conferência organizada pela SBC (Sociedade Brasileira de Computação) e afiliada, desde 2012, à ISCB (*International Society for Computational Biology*). Ela foi classificada pela CAPES como Qualis B4 em 2012.
- *International Conference on Bioinformatics and Computational Biology*: Conferência patrocinada pela ISCA (*International Society for Computers and their Applications*). Ela ainda não foi classificada pela CAPES.

Foram 8 publicações ao todo, tal como pode ser visto nas Referências. Cabe destacar que o autor da dissertação é o primeiro autor de todas as publicações.

3.2. Programas de Domínio Público

Além da ferramenta de auditoria, nosso trabalho gerou outros dois programas de domínio público como subproduto. Abaixo, apresentamos brevemente cada um deles, fornecendo informações sobre o endereço de acesso e a utilização por outros autores.

Nome: allPermutations
URL: `http://mirza.ic.unicamp.br:8080/bioinfo/download/allPermutations.zip`
Utilização por Outros Autores: [Labarre 2012]

Nome: Base de Dados de Distâncias de Rearranjo
URL: `http://mirza.ic.unicamp.br:8080`
Utilização por Outros Autores: [Grusea and Labarre 2011]

4. Conclusão

Neste trabalho, desenvolvemos uma ferramenta para avaliar as respostas de algoritmos de rearranjo de genomas. A fim de ilustrar sua aplicação, nós a utilizamos para avaliar as respostas de 16 algoritmos de rearranjo de genomas aproximados relativos a 6 variações do Problema da Ordenação por Rearranjo, dos quais 12 foram propostos na literatura e 4 foram propostos por nós.

Além da ferramenta, este trabalho trouxe outras contribuições. Desenvolvemos um algoritmo exato para calcular distâncias de rearranjo que é mais eficiente em termos de uso de memória do que qualquer outro algoritmo que encontramos na literatura. Apresentamos conjecturas que dizem respeito à forma como as distâncias de rearranjo se distribuem. Validamos conjecturas referentes ao diâmetro, que é o maior valor alcançável pela distância de rearranjo entre uma permutação qualquer e a identidade considerando-se todas as permutações com o mesmo número de elementos. Apresentamos demonstrações formais para o fator de aproximação de alguns dos algoritmos avaliados. Por fim, mostramos que o fator de aproximação de 7 dos 16 algoritmos avaliados não podem ser melhorados, o que contradiz algumas hipóteses levantadas na literatura, e conjecturamos que o fator de aproximação de outros 6 algoritmos também não podem.

Referências

Benoît-Gagné, M. and Hamel, S. (2007). A new and faster method of sorting by transpositions. In *Proceedings of the 18th Annual Symposium on Combinatorial Pattern Matching*, volume 4580 of *LNCS*, pages 131–141, London, ON, Canada. Springer-Verlag.

Dias, Z. and Meidanis, J. (2002). Sorting by prefix transpositions. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval*, volume 2476 of *LNCS*, pages 65–76, Lisbon, Portugal. Springer-Verlag.

Fischer, J. and Ginzinger, S. W. (2005). A 2-approximation algorithm for sorting by prefix reversals. In *Proceedings of the 13th Annual European Symposium on Algorithms*, volume 3669 of *LNCS*, pages 415–425, Mallorca, Spain. Springer-Verlag.

Galvão, G. R. and Dias, Z. (2011a). Computing rearrangement distace of every permutation in the symmetric group. In *Proceedings of the 26th ACM Symposium on Applied Computing – Conference Track on Bioinformatics and Computational Systems Biology*, pages 106–107, Taichung, Taiwan. ACM Press.

Galvão, G. R. and Dias, Z. (2011b). A flexible framework for computing rearrangement distance of every permutation in the symmetric group. In *Proceedings of the 6th Brazilian Symposium on Bioinformatics*, pages 33–40, Brasília, Brazil.

Galvão, G. R. and Dias, Z. (2011c). On the distribution of rearrangement distances. In *Proceedings of the 6th Brazilian Symposium on Bioinformatics*, pages 41–48, Brasília, Brazil.

Galvão, G. R. and Dias, Z. (2011d). On the performance of sorting by transpositions without using cycle graph. In *Proceedings of the 6th Brazilian Symposium on Bioinformatics*, pages 69–72, Brasília, Brazil.

Galvão, G. R. and Dias, Z. (2012a). GRAAu: Genome Rearrangement Algorithm Auditor. In *Proceedings of the 4th International Conference on Bioinformatics and Computational Biology*, pages 97–101, Las Vegas, NV, USA. Curran Associates, Inc.

Galvão, G. R. and Dias, Z. (2012b). On the approximation ratio for sorting by short swaps. In *Proceedings of the 7th Brazilian Symposium on Bioinformatics*, pages 120–125, Campo Grande, MS, Brazil.

Galvão, G. R. and Dias, Z. (2012c). On the approximation ratio of algorithms for sorting by transpositions without using cycle graphs. In *Proceedings of the 7th Brazilian Symposium on Bioinformatics*, volume 7049 of *LNCS*, pages 25–36, Campo Grande, MS, Brazil. Springer-Verlag.

Galvão, G. R. and Dias, Z. (2012d). On the performance of sorting permutations by prefix operations. In *Proceedings of the 4th International Conference on Bioinformatics and Computational Biology*, pages 102–107, Las Vegas, NV, USA. Curran Associates, Inc.

Grusea, S. and Labarre, A. (2011). The distribution of cycles in breakpoint graphs of signed permutations. *CoRR*, abs/1104.3353.

Kececioglu, J. D. and Sankoff, D. (1995). Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*, 13(1-2):80–110.

Labarre, A. (2012). Lower bounding edit distances between permutations. *CoRR*, abs/1201.0365.

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*, pages 96–102, Santa Cruz, Bolivia. IEEE Computer Society.

Walter, M. E. M. T., Dias, Z., and Meidanis, J. (2000). A new approach for approximating the transposition distance. In *Proceedings of the 7th International Symposium on String Processing Information Retrieval*, pages 199–208, Washington, DC, USA. IEEE Computer Society.