UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA

Uma implementação de Grafos De Bruijn a partir de Árvores de Sufixo

---

Proposta de Trabalho de Graduação

Aluno: João Lucas Gomes de Miranda
Orientador: Paulo Gustavo Soares da Fonseca

Sumário

1. Contexto

Atualmente, o processo de sequenciamento de DNA é efetuado principalmente utilizando-se as plataformas de sequenciamento de alto desempenho ditas de ``nova geração'' (*Next-Generation Sequencing---NGS*) [13] Essas tecnologias produzem um enorme volume de fragmentos curtos (comprimento abaixo das centenas) que precisam ser *montados*, i.e., alinhados e combinados, para reconstruir sequências originais de bilhões de letras.

As ferramentas para montagem de fragmentos NGS [6, 18, 4, 16] são majoritariamente baseadas nos chamados *Grafos de de Bruijn* (GDB). No GDB de ordem *k* construído a partir do conjunto de fragmentos *S, G(S),* os nós correspondem às subsequências de comprimento *k* (*k*-mers) das cadeias em S, e dois *k*-mers (nós) são unidos por uma aresta desde que haja uma sobreposição de tamanho *k-1*, de forma que as arestas correspondem aos *k+1*-mers de *S*. Efetuar a montagem de fragmentos usando GDB envolve problemas como o de encontrar *Caminhos Eulerianos*, que admite solução em tempo polinomial. Entretanto, um dos principais limitadores quanto ao emprego dessas técnicas é o espaço de memória exigido pelos GDB que, se representado explicitamente, pode requerer centenas de GigaBytes [9]. Diante disto, diversos esforços vêm sendo empreendidos para desenvolver estruturas de dados eficientes do ponto de vista de espaço, permitindo, todavia, operações sobre o GDB em tempo comparável a uma representação tradicional.

A representação e manipulação de GDB envolvem técnicas algorítmicas específicas que visam a minimizar o uso de memória principal. Essas técnicas incluem o desenho de algoritmos otimizados para memória externa/cache, o desenvolvimento de *estruturas de dados sucintas* [10], e o uso de compressão de dados.
Os métodos descritos na literatura procuram representar ou o conjunto de vértices (*k*-mers) ou arestas (*K+1*-mers) com estruturas de dados específicas de forma a permitir que as operações de navegação como ‘quais são os sucessores/predecessores de um nó *v*?’ sejam respondidas indiretamente através de operações primitivas de baixo nível. As estruturas de dados básicas incluem bit arrays comprimidos [9], Filtros de Bloom (+hash tables) [12, 7, 15], versões modificadas da Transformada de Burrows-Wheeler (+Wavelet Trees) [3, 2, 1] FM-Index [14, 8], e índices clássicos como árvores e arrays de sufixos [5].

2. Objetivo

Apesar do desenvolvimento das estruturas de dados para GDB ter-se dado historicamente de maneira *ad hoc*, respondendo a objetivos particulares de cada trabalho, partindo de representações muito básicas como hash tables, bit arrays e Filtros de Bloom, houve uma certa convergência na direção da associação entre os GDB e estruturas de índices e, em particular, por questões de eficiência, índices sucintos.

Os principais objetivos específicos deste trabalho de graduação são, portanto:

1. Estudar profundamente, do ponto de vista teórico, a associação entre GDB e índices de alto nível, em particular estruturas de índices como as árvores e arrays de sufixos, generalizando as formulações de [5];
2. Desenvolver implementações sucintas eficientes de GDB baseadas especificamente em Árvores de Sufixos - utilizando-se como ponto de partida as implementações em [19].

3. Cronograma

Atividade	**Março**		**Abril**		**Maio**		**Junho**	
Elaboração da Proposta	x	x						
Pesquisa		x	x	x	x	x		
Implementação				x	x	x	x	
Elaboração do Relatório					x	x	x	x
Preparação da Apresentação							x	x

Tabela 1: cronograma de atividades

4. Referências

[1] Djamal Belazzougui et al. “Bidirectional variable-order de Bruijn Graphs”. Em: LATIN 2016: Theoretical Informatics, 12th Latin American Symposium. 2016, pp. 164–178. ISBN: 978-3-662-49528-5. DOI: 10.1007/978-3-662-49529-2. URL: http://link.springer.com/10.1007/978-3-662-49529-2.

[2] Christina Boucher et al. “Variable-Order de Bruijn Graphs”. Em: Proceedings of the 2015 Data Compression Conference - DCC 2015. Snowbird, 2015, pp. 383–392. ISBN: 9781479984305. DOI: 10.1109/DCC. 2015.70. arXiv: 1411.2718.

[3] Alexander Bowe et al. “Succinct de Bruijn graphs”. Em: Proceedings of the Workshop on Algorithms in Bioinformatics -WABI 2012. Ljubljana: Springer, Berlin, Heidelberg, 2012, pp. 225–235. ISBN: 9783642331213. DOI: 10.1007/978-3-642-33122-0_18. URL: http://link.springer.com/10.1007/978-3-642-33122-0%7B%5C_%7D18.

[4] Jonathan Butler et al. “ALLPATHS: De novo assembly of whole-genome shotgun microreads”. Em: Genome Research 18.5 (2008), pp. 810–820. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.7337908. URL: http://genome.cshlp.org/content/18/5/810.

[5] Bastien Cazaux, Thierry Lecroq e Eric Rivals. “Linking indexing data structures to de Bruijn graphs: Construction and update”. Em: Journal of Computer and System Sciences In press (jul. de 2016). ISSN: 00220000. DOI: 10.1016/j.jcss.2016.06.008. URL: http://linkinghub.elsevier.com/retrieve/pii/S0022000016300502.

[6] Mark J Chaisson e Pavel A Pevzner. “Short read fragment assembly of bacterial genomes”. Em: Genome Research 18.2 (2008), pp. 324–330. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.7088808. URL: http://genome.cshlp.org/content/18/2/324.

[7] Rayan Chikhi e Guillaume Rizk. “Space-efficient and exact de Bruijn graph representation based on a Bloom Filter”. Em: Algorithms for Molecular Biology 8 (2013), p. 22. ISSN: 03029743. DOI: 10.1007/978-3-642-33122-0_19.

[8] Rayan Chikhi et al. “On the Representation of De Bruijn Graphs”. Em: Journal of Computational Biology 22.5 (maio de 2015), pp. 336–352. ISSN: 1066-5277. DOI: 10.1089/cmb.2014.0160. arXiv: 1401.5383. URL: http://online.liebertpub.com/doi/10.1089/cmb.2014.0160.

[9] Thomas C Conway e Andrew J Bromage. “Succinct data structures for assembling large genomes”. Em: Bioinformatics 27.4 (fev. de 2011), pp. 479–486. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq697. arXiv: 1008.2555. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq697.

[10] Guy Jacobson. “Succinct static data structures”. Tese de doutorado. Carnegie Mellon University, 1989. [11] Moritz Maaß. “Linear bidirectional on-line construction of affix trees”. Em: Algorithmica 37.1 (2003), pp. 43–74.

[12] Jason Pell et al. “Scaling metagenome sequence assembly with probabilistic de Bruijn graphs”. Em: Proceedings of the National Academy of Sciences 109.33 (2012), pp. 13272–13277. ISSN: 0027-8424.
DOI: 10.1073/pnas.1121464109. arXiv: arXiv:1112.4193v2.

[13] Mihai Pop e Steven L Salzberg. “Bioinformatics challenges of new sequencing technology”. Em: Trendsin Genetics 24.3 (2008), pp. 142–149. ISSN: 0168-9525. DOI: 10.1016/j.tig.2007.12.006. URL:
http://www.sciencedirect.com/science/article/pii/S016895250800022X.

[14] Einar Andreas Rødland. “Compact representation of k-mer de Bruijn graphs for genome read assembly.” Em: BMC bioinformatics 14.1 (2013), p. 313. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-313. URL:
http://www.biomedcentral.com/1471-2105/14/313.

[15] Kamil Salikhov, Gustavo Sacomoto e Gregory Kucherov. “Using cascading Bloom filters to improve the memory usage for de Brujin graphs”. Em: Algorithms for Molecular Biology 9 (fev. de 2014), p. 2. arXiv: 1302.7278. URL: http://arxiv.org/abs/1302.7278.

[16] Jared T Simpson et al. “ABySS: A parallel assembler for short read sequence data”. Em: Genome Research 19 (2009), pp. 1117–1123. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.089532.108. URL: http://genome.cshlp.org/content/19/6/1117.

[17] Dirk Strothmann. “The affix array data structure and its applications to RNA secondary structure analysis”. Em: Theoretical Computer Science 389.1-2 (2007), pp. 278–294. ISSN: 03043975. DOI: 10.1016/j.tcs.2007.09.029.

[18] Daniel R Zerbino e Ewan Birney. “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs”. Em: Genome Research 18.5 (fev. de 2008), pp. 821–829. ISSN: 1088-9051. DOI: 10.1101/gr. 074492.107. URL:
http://genome.cshlp.org/content/18/5/821%20http://www.genome.org/cgi/doi/10.1101/gr.074492.107.

[19] Simon Gog. “Succinct Data Structure Library 2.0”. URL:
https://github.com/simongog/sdsl-lite

5. Assinaturas

Recife, 12 de Abril de 2017.

_____

Paulo Gustavo Soares da Fonseca
(Orientador)

_____

João Lucas Gomes de Miranda
(Aluno)