Expansão do Espaço de Procura em Mapeamento Genómico de Alto Rendimento*

Natacha P. Leitão[1], João Leitão[2], and Francisco M. Couto[1]

[1] LaSIGE, Departamento de Informática,
Faculdade de Ciências, Universidade de Lisboa
[2] CITI, Departamento de Informática,
Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa
nleitao@lasige.di.fc.ul.pt jc.leitao@fct.unl.pt fcouto@di.fc.ul.pt

A história da sequenciação do DNA começou em 1977 com Sanger [1] e os seus trabalhos para determinar a ordem dos nucleótidos na sequência de DNA do bacteriófago phiX174; contudo, foi no início do século XXI que a tecnologia de sequenciação conheceu o seu maior desenvolvimento. Impulsionada pela sequenciação do genoma humano [2, 3], o desenvolvimento da nova geração de tecnologias de sequenciação (NGS) foi crucial no aumento da quantidade de dados de sequenciação (*reads*) e na diminuição do custo da sequenciação de DNA e de outros métodos da genómica [4]. Assim, a facilidade em gerar informação genómica contribui para tornar as NGS essenciais em várias áreas de investigação, como a Biologia Molecular e a Biologia Evolutiva, onde muitas vezes o primeiro passo consiste em mapear as *reads*, ou seja, alinhá-las a um genoma de referência. Existe, então, a necessidade de uma ferramenta de mapeamento rápida, precisa (que mapeie as *reads* nos locais certos do genoma de referência) e com elevada cobertura (percentagem de *reads* que são de facto mapeadas), que permita fazer face ao crescimento exponencial dos dados genómicos e facilite a sua interpretação biológica. Sobretudo porque as tecnologias atuais não são perfeitas e as *reads* contêm erros, como ter bases erradas em determinadas posições, um dos desafios levantados é a distinção entre os erros técnicos e as variações genéticas que ocorrem na amostra sequenciada.

Várias ferramentas têm sido propostas nos últimos anos [5] com diferentes abordagens, por exemplo, as que têm algoritmos de mapeamento 1) baseados em tabelas *hash* ou *seed-and-extend* (semelhante ao que ocorre no BLAST [6]), como o MAQ [7], o SeqMaq [8], o SHRiMP [9] e o GNUMAP [10]; ou 2) baseados em array de sufixos e compressão de genomas que utilizam o método de Burrows-Wheeler (BWT, do inglês *Burrows-Wheeler Transform*), como o BWA [11] e o Bowtie [12]; contudo, não cumprem os requisitos na totalidade. Os algoritmos do primeiro grupo, embora mais lentos, permitem um mapeamento mais preciso e uma maior cobertura, ou seja, mais *reads* são mapeadas e nos locais exatos; o último grupo, prima pela rapidez do processo, comprometendo a precisão e cobertura, sobretudo quando as diferenças entre a *read* e a sequência de referência são muitas (por exemplo, no mapeamento de *reads* de outras espécies nos estudos de comparação).

Neste artigo, é apresentada uma abordagem de expansão do espaço de procura sobre o genoma de referência, normalmente evitada por exigir uma capacidade de processamento e memória que nem sempre está disponível nas máquinas de quem vai fazer o mapeamento. À semelhança de ferramentas como o GNUMAP, inicialmente o algoritmo constrói uma tabela de *hash* para o genoma com chaves num determinado tamanho k, que servirão para localizar as *reads* no genoma. O que aqui se propõe é a criação de múltiplas chaves (de tamanho k) para cada *read* considerando que cada posição pode ter uma das quatro bases. Assim, apesar dos erros de sequenciação numa dada posição inicial, a *read*

* Este trabalho foi financiado pela Fundação para a Ciência e Tecnologia (FCT/MCTES) no contexto do Projecto Estratégico do LaSIGE, ref. PEst-OE/EEI/UI0408/2014.

terá várias chaves associadas a si mesma na procura da sua localização no genoma e a probabilidade de se encontrar a localização certa é maior; o mapeamento só será vinculativo após o alinhamento (utilizando o algoritmo probabilístico de Needleman-Wunsch [13]) positivo do resto da *read* com o genoma de referência. Considerando que, na generalidade dos casos, as *reads* apresentam melhores valores de qualidade no seu início, a explosão das chaves pode ficar restrita às posições com qualidade abaixo de um certo valor (por exemplo, 50% da base estar certa), reduzindo o número de chaves associadas. As vantagens da expansão do espaço de pesquisa para cada *read* será analisada recorrendo a *datasets* de NGS produzidos a partir do genoma de mamíferos (por exemplo, do *M. Muluscus*) e de *datasets* reais, de onde irão ser copiados os valores qualidade nos quais se vão basear os erros de sequenciação simulados. Os *outputs* serão, então, avaliados quanto ao número de *reads* não mapeadas, mapeadas em locais incorretos e mapeadas em mais do que um local. Um estudo cuidado será feito para se determinar o tamanho ideal para a chave das *reads* que permita fazer o balanço entre as elevadas precisão e cobertura, sem que o número de *reads* mapeadas a mais do que um local seja muito elevado. Por fim, uma vez que, um dos objetivos é desenhar o algoritmo para que seja altamente paralelizável podendo ser executado numa plataforma de *cloud computing*, torna-se uma ferramenta universal, que não está restringida pelas características da máquina do utilizador. Por outro lado, ao permitir ser executado entre várias máquinas a paralelização permite que o processo seja relativamente rápido, apesar de exigir maior processamento de informação.

Referências

1. Sanger, F., Nicklen, S., Coulson, A.R.: Dna sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences **74** (1977) 5463–5467
2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al.: The sequence of the human genome. science **291** (2001) 1304–1351
3. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.: Initial sequencing and analysis of the human genome. Nature **409** (2001) 860–921
4. Green, E.D., Guyer, M.S., Institute, N.H.G.R., et al.: Charting a course for genomic medicine from base pairs to bedside. Nature **470** (2011) 204–213
5. Fonseca, N.A., Rung, J., Brazma, A., Marioni, J.C.: Tools for mapping high-throughput sequencing data. Bioinformatics (2012) bts605
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215** (1990) 403–410
7. Li, H., Ruan, J., Durbin, R.: Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome research **18** (2008) 1851–1858
8. Jiang, H., Wong, W.H.: Seqmap: mapping massive amount of oligonucleotides to the genome. Bioinformatics **24** (2008) 2395–2396
9. Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A., Brudno, M.: Shrimp: accurate mapping of short color-space reads. PLoS computational biology **5** (2009) e1000386
10. Clement, N.L., Snell, Q., Clement, M.J., Hollenhorst, P.C., Purwar, J., Graves, B.J., Cairns, B.R., Johnson, W.E.: The gnumap algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. Bioinformatics **26** (2010) 38–45
11. Li, H., Durbin, R.: Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics **25** (2009) 1754–1760
12. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., et al.: Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol **10** (2009) R25
13. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology **48** (1970) 443–453