Otimização de desempenho de software para análises de diversidade genética utilizando programação paralela

Alexandre H. Aono, Álvaro L. Fazenda

[1]Instituto de Ciência e Tecnologia - Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brazil, 12231-280

alexandre.aono@gmail.com, alvaro.fazenda@unifesp.br

Abstract. *The use of molecular markers together with the main methodologies of genetic diversity allow us to infer how close organisms are on an evolutionary scale. In the present work, performance optimizations were performed in the calculation of genetic distance measurements in a R package, widely used by the scientific community with different data formats. With a mean reduction of 90 % in the time spent for these calculations using different measures of diversity, the results showed great optimization potential in methods of an area that is constantly evolving and has produced a massive amount of biological data.*

Resumo. *A utilização de marcadores moleculares em conjunto com as principais metodologias de diversidade genética permitem inferir o quão próximo determinados organismos se encontram em uma escala evolutiva. No presente trabalho, foram realizadas otimizações de desempenho no cálculo de medidas de distância genética em um pacote do R, amplamente utilizado pela comunidade científica com diferentes formatos de dados. Com uma redução média de 90% no tempo gasto para esses cálculos usando diferentes medidas de diversidade, os resultados mostraram grande potencial de otimização em métodos de uma área que se encontra em constante evolução e que tem produzido uma quantidade massiva de dados biológicos.*

1. Introdução

Toda a informação genética de um organismo é encontrada em seu genoma, que possui toda a informação hereditária de um ser codificada no seu DNA. O armazenamento de todos os dados transmitidos entre diferentes gerações é feito no genoma, cuja constituição vincula grupos de nucleotídeos a aminoácidos, formando proteínas com diferentes funções. Organismos de uma mesma espécie diferem em sequências de DNA e a existência de uma dada variação genética será influenciada por circunstâncias que os indivíduos passaram, incluindo o sucesso reprodutivo, migração, tamanho da população, seleção natural e eventos históricos [Sunnucks 2000]. Segundo [Mohammadi and Prasanna 2003], o estudo de diversidade genética é o processo pelo qual a variação entre indivíduos, grupos de indivíduos ou populações é analisado por um método específico ou uma combinação de métodos [Sunnucks 2000].

A mensuração da diversidade genética entre organismos é realizada por diferentes *softwares* biocomputacionais aplicados a dados genômicos. Esses programas são desenvolvidos com base em diferentes metodologias, modelos evolutivos, métricas e estatísticas. No presente trabalho, realizou-se a otimização de funções do *software* estatístico R para cálculo de distância genética em dados de marcadores moleculares, as

quais são utilizadas em diferentes etapas do processo de análise e representam uma etapa com custo computacional elevado no processo de estudo de diversidade.

2. Marcadores Moleculares e Diversidade Genética

Identificar os genótipos mais relevantes a uma dada característica é objetivo chave na Genética [Schlötterer 2004], sendo essas mudanças herdáveis, a matéria-prima para os processos de melhoramento genético e evolução [Ramalho et al. 1990]. Essa distinção geralmente é baseada em sistemas informativos de marcadores moleculares [Schlötterer 2004], os quais são vistos como ferramentas genéticas que nos permitem examinar diferenças entre indivíduos em diversas posições do genoma. De acordo com [Griffiths et al. 2006], marcadores moleculares são regiões genômicas de heterozigose molecular, ou seja, sequências de nucleotídeos que podem ser investigadas por intermédio dos polimorfismos presentes em diferentes indivíduos, ocasionados por inserções, deleções, eventos de duplicação, translocação, etc. [Nadeem et al. 2018].

Os marcadores moleculares mais utilizados em diferentes aplicações [Schlötterer 2004, Nadeem et al. 2018] são os polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms* - SNPs) e Microssatélites (*Single Sequence Repeats* - SSRs). SNPs são mutações de uma única base entre diferentes indivíduos de uma mesma espécie e são encontrados em grande volume pelo genoma [Griffiths et al. 2006]. O microssatélite é uma classe de sequências de DNA repetitivo que ocorre em todos os organismos e consiste em repetições de sequências geralmente de dois a seis nucletotídeos, ocupando uma extensão de até 100 pares de bases [Ramalho et al. 1990]. Um marcador SSR é baseado no número dessas repetições ao longo de uma região cromossômica [Griffiths et al. 2006].

Após a obtenção dos marcadores, a análise dos dados geralmente envolve medições numéricas e, em muitos casos, combinações de diferentes tipos de variáveis [Mohammadi and Prasanna 2003]. Nesse contexto, a utilização de distâncias genéticas ou medidas de distância tradicionais, como a Euclidiana, representam uma maneira de mensurar a diferenciação entre pares de observações. De forma simplificada, a mensuração dessas diferenciações ocorre percorrendo-se os n indivíduos e calculando-se as distâncias contra os $n - 1$ indivíduos restantes.

3. Métodos

Os métodos adaptados e otimizados para cálculo de distância genética foram baseados no pacote *poppr* [Kamvar et al. 2014] do software estatístico R. Foram utilizados os seguintes métodos para cálculo de distância: (a) Euclidiana; (b) Edwards; (c) Nei; (d) Prevosti; (e) Reynolds e (f) Rogers. Visando obter ganho no desempenho computacional, implementaram-se versões destes procedimentos utilizando a linguagem C. Esses códigos foram acoplados ao pacote original do software em R, em conjunto com a utilização de programação paralela por meio da biblioteca OpenMP, além do uso de funções matemáticas da bem conhecida biblioteca BLAS (*Basic Linear Algebra Subprograms*). Para análises de desempenho utilizou-se um nó de processamento composto por duas CPUs Intel Xeon E5-2660-v4@2.00GHz com 28 núcleos (*cores*).

Como conjuntos de dados utilizados para teste, realizou-se a construção de tabelas simuladas com marcadores microssatélites. De modo a representar situações diversas em

um cenário científico na área biológica, foram utilizadas as seguintes configurações de tamanho: (I) 100 indivíduos e 8.000 marcadores; (II) 100 indivíduos e 10.000 marcadores; (III) 200 indivíduos e 10.000 marcadores; (IV) 400 indivíduos e 10.000 marcadores. Para avaliação dos resultados obtidos com a otimização, foram utilizadas métricas de *Speed-up* e Eficiência com as diferentes configurações. Com o mesmo ambiente computacional para execução, foram testadas em 3 rodadas diferentes quantidades de *threads* para avaliação do desempenho do projeto de algoritmo paralelo (1, 2, 3, 4, 5, 10 *threads*). Além disso, mediu-se o tempo gasto para construção da matriz de distâncias usando as versões originais da função em R e suas respectivas modificações. A limitação de 10 *threads* ocorreu em função do foco do desenvolvimento residir em computadores pessoais típicos de pesquisadores da área.

4. Resultados

A primeira etapa no processo de otimização foi a criação de um código em C acoplado a R para comparação com o desempenho das funções tradicionais. Para o processo de paralelização, optou-se pela utilização de um escalonamento de iterações no laço entre as *threads* de forma não usual, conhecido como guiado (*schedule(guided)*), dado que havia desbalanceamento de carga em alguns procedimentos. Somente com a alteração de linguagem, houve um ganho significativo na diminuição do tempo nas funções para cálculo de distância Euclidiana, de Nei e de Prevosti (48; 86,72 e 91,91% respectivamente). Em relação às demais funções (Edwards, Nei e Reynolds), que se utilizam de cálculos matemáticos mais complexos, foi necessária a utilização de procedimentos disponíveis na biblioteca BLAS, representando um ganho de 91% na diminuição do tempo no cálculo da função de Nei, 90% na função de Edwards e 92% na função de Reynolds.

Para avaliação do desempenho paralelo, construiu-se um gráfico com a distribuição das medições de tempo normalizadas nos maiores tamanhos de entrada tomados para teste. A figura 1 mostra de maneira simplificada o decaimento de tempo do cálculo das distâncias com diferentes quantidades de *threads*. Conforme pode ser observado, os ganhos mais representativos foram com as funções implementadas sem a utilização do BLAS. O processo de paralelização com a multiplicação de matrizes pela biblioteca não é passível de modificação e, dessa forma, as maiores restrições ao ganho de desempenho são em virtude do funcionamento da biblioteca. A tabela 1 mostra os *Speedups* obtidos com 10 *threads* de duas distintas formas: *Speedup* paralelo tomando por base o tempo de execução com uma única *thread*, e *Speedup* tomando por base a versão original executada dentro do Software R.

Tabela 1. *Speedups* e Eficiência medidos com o código acelerado em relação a versão paralela (VP) com dez *threads* e versão original

	Medida de distância					
Medidas	Roger	Reynolds	Prevosti	Nei	Euclidiana	Edwards
Speedup Original	92,41	80,49	64,08	45,92	18,4	97,58
Speedup VP	9,3	6,4	8,7	3,7	8,3	7,8
Eficiência VP	93,10%	64,24%	87,41%	37,78%	83,24%	78,79%

5. Conclusão

As aplicações de métodos para cálculo de distância genética são imensuráveis, desde a utilização das distâncias em conjunto com técnicas multivariadas a análises individuais

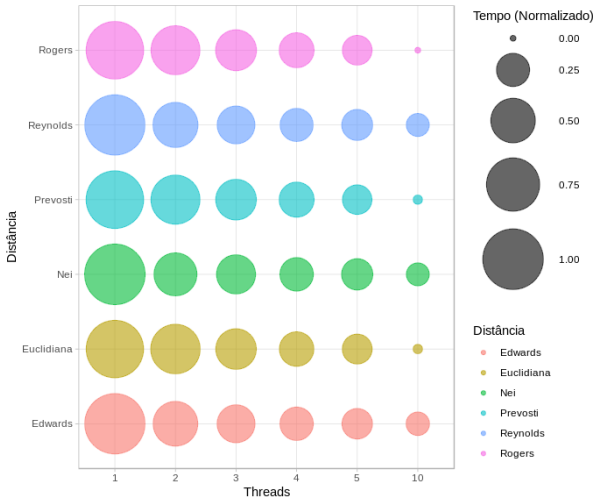

Figura 1. Comparação de tempo com valores de dimensão máximos.

para mensuração de diferenciações entre subgrupos ou organismos específicos. Devido ao custo computacional elevado, a utilização do software estatístico R e ao potencial volume de dados em tais análises, os cálculos de distância representam uma etapa crítica em um processo de exploração da diversidade genética em populações. Os resultados obtidos demonstram a vasta gama de possibilidades de ganho de desempenho em métodos de análise genética e consequentemente maior celeridade na pesquisa e aproveitamento de recursos computacionais. As melhorias atingidas no processamento estão acopladas a um sistema que está em desenvolvimento para realização automatizada de análises de diversidade genética. Esta ferramenta será disponibilizada em publicação futura.

Referências

Griffiths, A. J., Wessler, S. R., Lewontin, R. C., Gelbart, W. M., Suzuki, D. T., and Miller, J. H. (2006). Introdução à genética. In *Introdução à genética*.

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). Poppr: an r package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2:e281.

Mohammadi, S. and Prasanna, B. (2003). Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop science*, 43(4):1235–1248.

Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., Hatipoğlu, R., Ahmad, F., Alsaleh, A., Labhane, N., et al. (2018). Dna molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnology & Biotechnological Equipment*, 32(2):261–285.

Ramalho, M., dos Santos, J. B., and Pinto, C. B. (1990). *Genética na agropecuária*. FAEPE.

Schlötterer, C. (2004). The evolution of molecular markers—just a matter of fashion? *Nature reviews genetics*, 5(1):63.

Sunnucks, P. (2000). Efficient genetic markers for population biology. *Trends in ecology & evolution*, 15(5):199–203.