POTION: um software paralelizado para a detecção de grupos de genes homólogos sob evidência de seleção positiva em escala genômica

Jorge Augusto Hongo[1]
Francisco Pereira Lobo[2]

Introdução

Uma fração considerável dos genes encontrados em projetos genoma não possui função biológica conhecida (REICHARDT, 2007). Esse vasto universo de genes desconhecidos constitui um campo fértil para a busca de genes interessantes, visando aplicações de biotecnologia. No caso de espécies de interesse agropecuário, esses genes desconhecidos constituem um vasto campo de buscas para localização de genes de interesse para ganhos de produção (CORBI et al., 2011; FAN et al., 2009; GU et al., 2009). Nesse cenário, é de extrema importância que novos métodos computacionais sejam desenvolvidos para a detecção de genes desconhecidos que apresentem potencial de contribuir para traços fenotípicos interessantes em espécies animais e vegetais estudadas pela Embrapa. Uma estratégia ainda não explorada para a detecção de genes potencialmente interessantes é a busca por grupos de genes homólogos (grupos de genes encontrados em espécies diferentes) sob evidência de seleção positiva (AGUILETA et al., 2009). A vasta maioria dos genes homólogos é conservada. Isso ocorre porque usualmente mutações não-sinônimas diminuem a eficiência funcional da proteína, o que diminui a aptidão evolutiva do indivíduo e impede a fixação do novo alelo quando comparado ao alelo ancestral (HARTWELL, 2011). Entretanto, alguns poucos grupos de

[1] Unicamp, Instituto de Computação, giovanni.castro@colaborador.embrapa.br
[2] Embrapa Informática Agropecuária, francisco.lobo@embrapa.br

genes homólogos evoluem apresentando uma forte pressão seletiva para a variação, ao invés da conservação (YANG, 2007). Uma vez que as espécies estudadas pela Embrapa têm sido alvo de seleção artificial para alguns poucos fenótipos de interesse visando ganho de produtividade, é razoável supor que os genes sob evidência de seleção positiva nessas espécies serão, possivelmente, associados a fenótipos de produtividade (CORBI et al., 2011; FAN et al., 2009; GU et al., 2009). Nesse contexto, a busca por genes sob evidência de seleção positiva em genomas de espécies de interesse da Embrapa constitui uma importante ferramenta para indicar possíveis genes associados a um maior ganho de produção nessas espécies. Entretanto, diversos dos passos para a detecção de seleção positiva são computacionalmente custosos. Para contornar tal problema, uma possível estratégia seria o desenvolvimento de progamas paralelizados, uma vez que a detecção de seleção positiva em cada grupo de homólogos independe das buscas realizadas em outros grupos. O presente trabalho descreve o desenvolvimento do software POTION (POsitive selecTION) para a busca por grupos de genes homólogos sob evidência de seleção positiva.

Material e métodos

POTION é um software modular e facilmente expansível que utiliza diversos programas que são o estado-da-arte em seus respectivos campos, tais como OrthoMCL para a detecção dos grupos de homólogos (CHEN et al., 2006), MUSCLE para o alinhamento dos grupos de proteínas homólogas (EDGAR, 2004; RETIEF, 2000) phylip para a construção de árvores filogenéticas (RETIEF, 2000) e PAML para a detecção de seleção positiva (YANG, 2007). O POTION é capaz de adequar os arquivos de saída de cada um dos software listados acima para o próximo software da *pipeline*. O programa final produzido possui aproximadamente 1500 linhas de código e utiliza diversos módulos sofisticados de bioinformática previamente desenvolvidos para perl (bioperl)[1]. O usuário pode controlar o comportamento de todos os softwares de terceiros por parâmetros globais definidos no início da execução da *pipeline*.

[1] Disponível em: <http://www.bioperl.org/wiki/Main_Page>.

Resultados e discussão

Para validar o POTION foi utilizado um padrão-ouro que consiste em 40 grupos de paralogos do parasita Trypanosoma brucei previamente investigados para a busca de seleção positiva, dos quais 23 apresentaram evidência de seleção positiva. O POTION foi capaz de detectar seleção positiva em 22 dos 23 genes previamente identificados, e classificou de maneira errônea somente um gene, o qual não possuia seleção positiva no padrão-ouro e foi classificado como positivamente selecionado pelo software. Dessa maneira, a ferramenta apresentou valores de sensibilidade, especificidade e F-measure de 0.96. O tempo total para analisar o padrão-ouro diminui de maneira proporcional ao número de processadores utilizados na análise, demonstrando que a paralelização do software é satisfatória. O elevado valor de F-measure observado, associado à parelização do POTION, demonstra que este software pode ser efetivamente adotado em uma ampla gama de estudos bioinformáticos onde a identificação de seleção positiva em escala genômica é um fator chave.

Agradecimentos

À Embrapa, por fornecer a insfraestrutura computacional para a realização deste trabalho.

Referências

AGUILETA, G.; REFREGIER, G.; YOCKTENG, R.; FOURNIER, E.; GIRAUD, T. “Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists.” **Infection, Genetics and Evolution**, v. 9, n. 4, p. 656-670, 2009.

CHEN, F.; MACKEY, A. J.; STOECKERT JUNIOR, C. J.; ROOS, D. S. “OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.” **Nucleic Acids Research**, v. 34 (Database issue), p. D363-368, Jan. 2006.

CORBI, J.; DEBIEU, M.; ROUSSELET, A.; MONTALENT, P.; LE GUILLOUX, M.; MANICACCI, D.; TENAILLON, M. I. “Contrasted patterns of selection since maize domestication on duplicated genes encoding a starch pathway enzyme.” **Theoretical and Applied Genetics**, v. 122, n. 4, p. 705-722, 2011.

EDGAR, R. C. “MUSCLE: a multiple sequence alignment method with reduced time and space complexity.” **BMC Bioinformatics**, v. 5, p. 113, 2004.

ESTEBAN, D. J.; HUTCHINSON, A. P. “Genes in the terminal regions of orthopoxvirus genomes experience adaptive molecular evolution.” **BMC Genomics**, v. 12, p. 26, 2011.

FAN, L.; BAO, J.; WANG, Y.; YAO, J.; GUI, Y.; HU, W.; J.; ZHU, J.; ZENG, M.; LI, Y.; XU, Y. “Post-domestication selection in the maize starch pathway.” **PLoS ONE**, v. 4, n. 10, p. e7612, 2009.

GU, J.; ORR, N.; PARK, S. D.; KATZ, L. M.; SULIMOVA, G.; MACHUGH, D. E.; HILL, E. W. “A genome scan for positive selection in thoroughbred horses.” **PLoS ONE**, v. 4, n. 6, p. e5767, 2009.

HARTWELL, L. **Genetics**: from genes to genomes. 4th ed. New York: McGraw-Hill, 2011. v. 1.

REICHARDT, J. K. “Quo vadis, genoma? A call to pipettes for biochemists.” **Trends Biochem SciENCE**, v. 32, n. 12, p. 529-530, 2007.

RETIEF, J. D. “Phylogenetic analysis using PHYLIP.” **Methods in Molecular Biology**, n. 132: 243-258, 2000.

YANG, Z. “PAML 4: phylogenetic analysis by maximum likelihood.” **Molecular Biology and Evolution**, v. 24, n. 8, p. 1586-1591, May, 2007.