Desenvolvimento de software para a detecção de grupos de genes homólogos sob evidência de seleção positiva

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

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 Empresa Brasileira de Pesquisa Agropecuária (Embrapa). Uma estratégia ainda não explorada pela Embrapa 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). É fato amplamente conhecido que 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 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 artifi-

[1] *UNICAMP, Instituto de Computação, jorgeahongo@gmail.com*
[2] *Embrapa Informática Agropecuária, francisco@cnptia.embrapa.br*

cial para alguns poucos fenótipos de interesse, visando à 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. Assim, o presente trabalho descreve o desenvolvimento de um software para a busca por grupos de genes homólogos sob evidência de seleção positiva. Nesse trabalho, desenvolveu-se um software em perl para integrar diversos softwares de terceiros capazes de realizar as tarefas individuais para a detecção de genes sob evidência de seleção positiva, conforme resumido na Figura 1. Nessa figura as caixas em azul indicam os arquivos produzidos sequencialmente durante a análise; as caixas em laranja indicam os módulos perl desenvolvidos (os quais controlam a execução coordenada dos softwares de terceiros utilizados); as setas em laranja indicam o fluxo sequencial da análise; as caixas em verde indicam softwares desenvolvidos por terceiros utilizados neste trabalho. O cilindro em azul indica o banco de dados desenvolvido para organizar e armazenar os dados produzidos. As letras sequenciais em vermelho indicam as etapas necessárias para que uma análise seja feita. Todas as etapas da análise

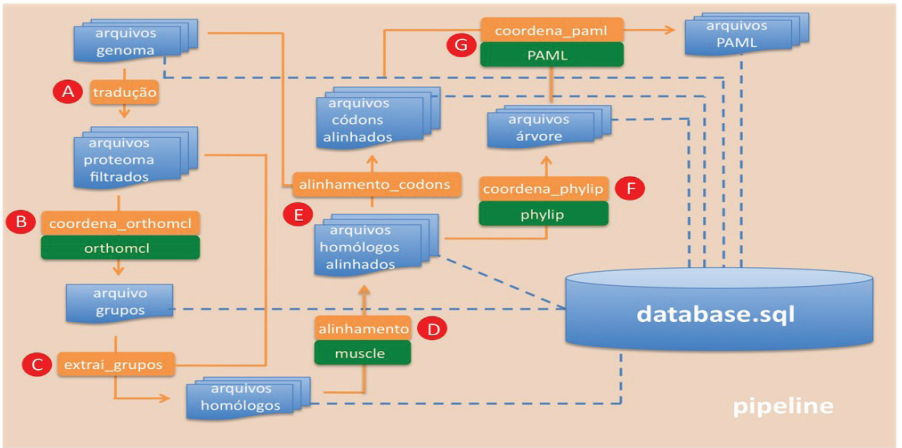

Figura 1. Arquitetura do sistema.

foram desenvolvidas de maneira modular, de modo a permitir o futuro uso de programas alternativos em cada uma das etapas. Para o desenvolvimento do software para a detecção de seleção positiva em grupos de genes homólogos, empregou-se os softwares OrthoMCL para a detecção dos grupos de homólogos (CHEN et al., 2006), MUSCLE para o alinhamento dos homólogos (EDGAR, 2004; RETIEF, 2000) phylip para a construção de árvores filogenéticas (RETIEF, 2000) e PAML para a localização de seleção positiva (YANG, 2007). Foram desenvolvidos procedimentos para adequar cada um dos arquivos de saída de cada um dos softwares listados acima para o próximo software da *pipeline*, conforme ilustrado na Figura 1 (caixas laranja). O programa final produzido possui aproximadamente 1000 linhas de código e utiliza diversos módulos sofisticados de bioinformática previamente desenvolvidos para perl (bioperl, http://www.bioperl.org/wiki/Main_Page). O usuário pode controlar o comportamento de todos os softwares de terceiros utilizados através de parâmetros globais definidos no início da execução da *pipeline*. Diversos conjuntos de dados na literatura científica são compostos por buscas manuais, por genes sob evidência de seleção positiva, e serão utilizados como provas de conceito do software desenvolvido (CORBI et al., 2011; ESTEBAN; HUTCHINSON, 2011; FAN et al., 2009).

Os autores agradecem à Embrapa por fornecer a bolsa, para que o estagiário desenvolvesse o trabalho em questão.

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.