



**DETECÇÃO E ANÁLISE BIOINFORMÁTICA DE GENES SOB EVIDÊNCIA DE
SELEÇÃO POSITIVA EM GENOMAS DE PARASITOS**

JORGE A. HONGO¹; ADHEMAR Z. NETO²; FRANCISCO P. LOBO³

Nº 12612

RESUMO

A relação ecológica de parasitismo é uma constante corrida armamentista entre os organismos parasitas e seus hospedeiros. A infecção por parasitas diminui a aptidão evolutiva dos hospedeiros e, conseqüentemente, mecanismos anti-parasitismo são positivamente selecionados continuamente dentre o conjunto de genes que compõem o genoma do organismo hospedeiro. Entretanto, a seleção positiva de mecanismos anti-parasitismo por parte dos hospedeiros impõe novas pressões seletivas aos organismos parasitas. Dessa maneira, genes de parasitas que permitam o escape dos mecanismos anti-parasitismo do hospedeiro aumentam a aptidão evolutiva do organismo parasita, sendo também selecionados positivamente. Esse fenômeno acaba por causar uma espiral de eventos coevolutivos ao longo do tempo em ambos os genomas no que se refere aos genes envolvidos na relação molecular parasito-hospedeiro. Genes evoluindo sob esse tipo de pressão seletiva no sistema parasita-hospedeiro muitas vezes apresentam uma freqüência de mutações não-sinônimas e sinônimas mais elevada do que a da vasta maioria dos outros genes destes genomas, fenômeno este denominado seleção positiva. Assim, dentre todos os genes observados no genoma de hospedeiros e parasitas, genes sob evidência de seleção positiva são ótimos candidatos a genes envolvidos na relação ecológica de parasitismo. Entretanto, o software existente para o cálculo de seleção positiva é computacionalmente custoso, tornando proibitivo a busca por seleção positiva em escala genômica. Nesse cenário, o presente trabalho descreve um software que faz uso de paralelização para permitir a busca por seleção positiva em escala genômica em tempo exequível.

1 Bolsista CNPq: Graduação em Eng. de Computação, UNICAMP, Campinas-SP, jorgeahongo@gmail.com.

2 Colaborador: Pesquisador, Embrapa Informática Agropecuária, Campinas-SP.

3 Orientador: Pesquisador, Embrapa Informática Agropecuária, Campinas-SP.



ABSTRACT

The biological interaction of parasitism is a constant arms race of hosts and parasites. The infection by parasites decreases the host fitness and, consequently, anti-parasitic mechanisms are positively selected from the genes that compose the host genome. On the other hand, the positive selection of anti-parasitic mechanisms on host genomes generate a selective pressure on parasite genomes, and parasite genes that allow them to escape the host anti-parasitic mechanisms increase the fitness of parasites, and are also positively selected. This phenomenon ends up in a spiral of coevolutionary events in time, where the genomes of both hosts and parasites are continuously positively selected on genes related to the host-parasite biological interaction. Genes evolving under such pressures usually present a much higher ratio of non-synonymous mutations when compared with the ratio of synonymous mutations, a phenomenon known as positive selection. Therefore, from all genes contained on host and parasite genomes, the ones presenting positive selection signatures are the ones more likely to be involved in host-parasite interactions. However, the software to detect positive selection are computationally time-costly, prohibiting the genomic-scale search for positive selection. In this scenario, the present work describes a software to detect positive selection that uses parallelization solutions in order to search for potentially interesting genes in genomic scale data in a feasible time.

INTRODUÇÃO

A ciência genômica é, hoje, uma realidade. Os dados de seqüenciamento de genomas completos encontram-se na ordem dos milhares, e muito mais genomas são esperados com a adoção generalizada das tecnologias de seqüenciamento de DNA de próxima geração (1,2). Uma fração considerável dos genomas já seqüenciados correspondem à organismos parasitas, imensamente estudados dado as grandes perdas que causam para a saúde humana, animal e vegetal. A disponibilidade de dados genômicos de parasitas tende a crescer substancialmente nos próximos anos, uma vez que um dos principais usos das tecnologias de seqüenciamento de DNA de próxima geração é o reseqüenciamento de múltiplas linhagens de organismos patogênicos já seqüenciados, permitindo assim a compreensão genética das diferenças fenotípicas entre as linhagens (3). Dentre os estudos pós-genômicos factíveis para organismos parasitas destaca-se a detecção e caracterização de regiões codificadoras sob evidência de seleção positiva visando localizar genes possivelmente envolvidos na relação parasita-hospedeiro (4,5). O desenvolvimento e a validação um software para a detecção computacional em escala genômica de genes sob evidência de seleção positiva em grupos de homólogos consistiu no objeto de estudo deste trabalho.

O estudo molecular da ação da seleção natural em regiões codificadoras evidencia um claro viés na freqüência de mutações não-sinônimas quando comparada à freqüência de mutações sinônimas (6). De maneira geral, alinhamentos múltiplos de códons de um dado grupo de genes homólogos possuem a vasta maioria das colunas do alinhamento sem variação no aminoácido codificado; já as mutações para códons sinônimos ocorrem em freqüências consideravelmente maiores. Este fenômeno ocorre porque mutações não-sinônimas usualmente reduzem a eficiência funcional da proteína codificada em comparação ao alelo não mutante fixado anteriormente. Assim, mutações não-sinônimas geralmente diminuem a aptidão evolutiva dos organismos que as possuem, e estes alelos menos funcionais são rapidamente removidos das populações através de seleção negativa ou purificadora (4,7). Entretanto, algumas poucas posições em alguns poucos genes podem apresentar uma freqüência de mutações não-sinônimas significativamente mais elevada do que o observado no restante dos alinhamentos de códons em análise, indicando a preferência pela fixação de novos alelos em detrimento aos antigos. Este fenômeno é denominado seleção positiva ou seleção Darwiniana, sendo observado em códons e genes que codificam

proteínas nas quais ocorre pressão seletiva para a variação ao invés da conservação do aminoácido na posição em análise quando comparada ao restante das posições, ou em uma dada seqüência quando comparada ao restante das seqüências. Dentre os fenômenos biológicos que comumente possuem grupos de genes homólogos evoluindo sob pressão seletiva positiva destacam-se genes envolvidos em percepção sensorial, reprodução, imunidade e na relação parasita-hospedeiro (4,5,8).

Espécies parasitas e hospedeiras encontram-se em uma contínua corrida armamentista: parasitas evoluindo para escapar do sistema imune hospedeiro e hospedeiros evoluindo para eliminar os parasitas através de seu sistema imune, conforme ilustrado pela amplamente aceita teoria da Rainha Vermelha (9). Parasitas dependem dos recursos do hospedeiro para a sua sobrevivência, e mutações que gerem o aumento de aptidão evolutiva em uma espécie parasita por favorecer características parasíticas são selecionadas positivamente. Entretanto, mutações que aumente a aptidão evolutiva de parasitas acabam por diminuir a aptidão evolutiva de indivíduos da espécie hospedeira infectadas pelos parasitas em questão, uma vez que tendem a aumentar a infectividade da espécie parasita. Conseqüentemente, mecanismos anti-parasitários que porventura surjam nas espécies hospedeiras aumentarão a aptidão evolutiva dos organismos hospedeiros que possuem tais mecanismos, sendo também selecionados positivamente. O ganho de aptidão evolutiva em uma das espécies desencadeia pressão seletiva para o ganho de aptidão na outra espécie envolvida na relação ecológica de parasitismo, desencadeando uma espiral de eventos co-evolutivos de seleção positiva em ambas as espécies. Assim, espera-se que diversos dos genes envolvidos na relação parasita-hospedeiro encontrem-se comumente sob forte seleção positiva em ambas as espécies envolvidas nessa relação ecológica (4,5). Conforme esperado, genes envolvidos na relação com os hospedeiros encontram-se sob forte pressão seletiva positiva nas mais diversas classes de parasitas, tais como vírus (10), bactérias (11), protozoários (12), insetos (13) e helmintos (14), dentre outros.

Diversos dos procedimentos computacionais parcialmente necessários para a localização de genes evoluindo sob seleção positiva (alinhamento de seqüências, construção de árvores filogenéticas e detecção de seleção positiva) encontram-se desenvolvidos. Entretanto, nenhuma abordagem ampla no sentido de se integrar estes diversos procedimentos computacionais em um único software para a detecção automatizada de regiões sob evidência de seleção positiva foi realizada. Um outro

agravante para a detecção em larga escala de grupos de genes homólogos sob evidência de seleção positiva é o tempo computacional gasto em cada análise. Estimativas de tempo para o cálculo de seleção positiva nos grupos de genes homólogos no genoma de vertebrados de são da ordem de 100 anos de computação (15). Uma vez que o resultado da busca por seleção positiva em cada grupo de genes homólogos não depende dos resultados das outras buscas, cada grupo de genes homólogos pode ser analisado em paralelo, de modo a minimizar o tempo total de computação. Nesse contexto, o presente projeto visa desenvolver procedimentos computacionais automatizados para a detecção de genes sob evidência de seleção positiva em genomas completos de parasitas, visando escolher novos alvos para a priorização de estudos pós-genômicos visando, em última análise, a priorização de alvos para o desenvolvimento de fármacos e vacinas.

MATERIAL E MÉTODOS

Desenvolvimento do software POTION

Conceito

As etapas bioinformáticas para a detecção de seleção positiva em grupos de genes homólogos encontra-se descrita na Figura 1, a qual será usada como guia nesta seção. Para a detecção de seleção positiva em escala genômica, três conjuntos de dados são necessários, para cada grupo de genes homólogos: Um arquivo de árvore filogenética descrevendo a relação filogenética entre as seqüências do grupo (Figura 1 F), um arquivo de alinhamento de códons (Figura 1 E), descrevendo o padrão de substituição entre as seqüências, e um arquivo que explicita quais são os modelos evolutivos a serem testados (Figura 1 G e H). Usualmente testa-se um modelo nulo, o qual calcula a probabilidade daquele alinhamento de códons ser derivado daquela árvore filogenética assumindo que ocorrem somente dois tipos de sítios de códons, alvos de seleção negativa ou neutra. Este modelo é comparado a um modelo alternativo, o qual calcula a mesma probabilidade do modelo anterior admitindo a existência também de uma terceira classe de sítios, os alvo de seleção positiva. Caso o modelo alternativo apresente uma maior probabilidade de ocorrência do que o modelo nulo (através de um teste de verossimilhança implementado

conforme sugerido por (16,17)) o grupo de genes homólogos é dito sob evidência de seleção positiva.

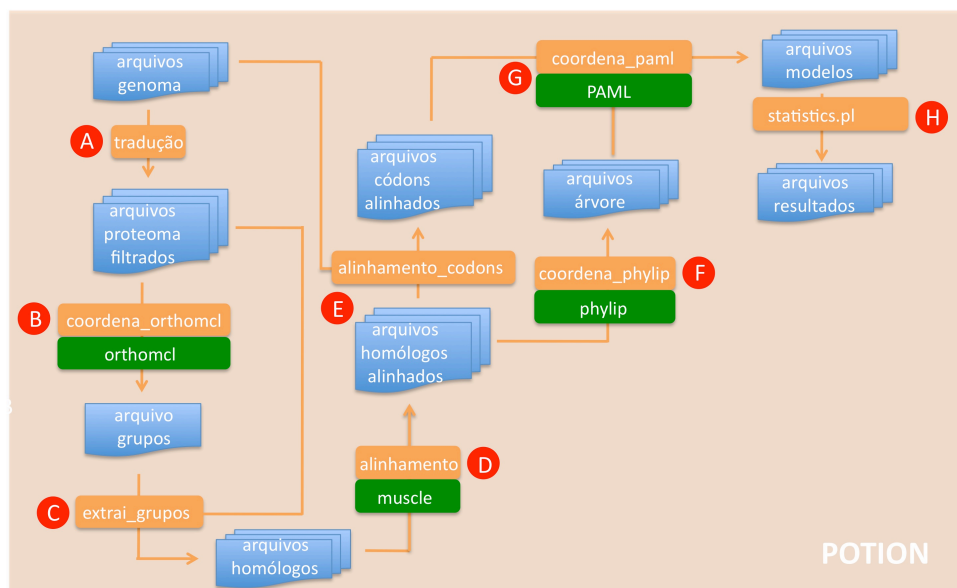


Figura 1. Esquema do software POTION. Retângulos azuis indicam arquivos produzidos, laranja indicam módulos de POTION, verdes indicam software pré-existent. A – tradução automática das seqüências codificadoras extraídas dos genomas; B – predição dos grupos de homólogos a partir das seqüências protéicas; C – definição dos arquivos de grupos de homólogos de proteínas; D – alinhamento múltiplo das seqüências homólogas protéicas; E – uso do alinhamento protéico para realizar o alinhamento de códons; F – cálculo das árvores filogenéticas; G – uso do alinhamento de códons e da árvore filogenética para o cálculo dos modelos nulo e alternativo no que se refere à presença de seleção positiva; H – cálculo dos valores p- e q- para significância.

Implementação

O programa POTION utiliza diversos programas estado da arte para diversas etapas, tais como MUSCLE para alinhamento múltiplo de proteínas (18), phylip para a construção de árvores filogenéticas (19) e PAML para o cálculo de verossimilhança de modelos de substituição de códons (17), representados em verde na Figura 1. Em laranja observa-se os diferentes módulos desenvolvidos por nosso grupo visando controlar e executar sequencialmente os passos necessários para a detecção em escala genômica de seleção positiva. Em azul observa-se os arquivos intermediários produzidos durante a análise. Em laranja observa-se os módulos de POTION que encapsulam os programas externos (em verde) ou executam tarefas internas do programa. O programa POTION recebe como parâmetros de entrada arquivos de

genoma contendo genes preditos (formato genbank) e um arquivo de configurações que contém virtualmente todos os parâmetros de configuração dos software que o programa POTION executa. Para reduzir o tempo de processamento, o programa paraleliza a análise dos grupos de ortólogos identificados entre os processadores disponibilizados pelo usuário (Figura 1 etapas D, E e F). Outro nível de otimização envolveu a paralelização das as instâncias do PAML para os quatros modelos utilizados na análise, com implementação de camadas de transição entre as tarefas para gerenciar os processadores alocados, necessário para garantir que nenhum processador esteja ocioso em qualquer momento que haja uma tarefa pronta para ser realizada (Figura 1, etapa G). O gerenciador também permite priorizar quais modelos do PAML devem ser executados e utiliza como heurística de priorização a alocação dos modelos com maior tempo estimado de processamento para garantir a alocação dos menos demorados para os demais processadores em concorrência, estabelecendo que o tempo de processamento do programa seja limitado ao do modelo mais longo para um número suficiente de processadores ao invés de uma combinação do tempo dos modelos utilizados pelo PAML. O programa também passou por vários processos de refatoração ao longo do seu desenvolvimento para remover desperdícios no uso dos recursos alocados ao programa e para garantir a facilidade de manutenção do código.

Conjunto de dados de teste

Utilizamos como conjunto de teste uma análise genômica de seleção positiva previamente realizada utilizando 101 vírus do grupo orthopoxvirus (20). Estes vírus causam doenças no homem e em outros animais, como a varíola e a vaccinia bovina, e o número de genes destes genomas varia de 70 a cerca de 100 genes. A existência de um conjunto de genomas pequenos já estudado em termos de seleção positiva permitiu que conseguíssemos avaliar continuamente o desempenho das diversas versões do software POTION em termos do ganho de desempenho computacional e em termos da detecção de conjuntos de genes homólogos sabidamente sob evidência de seleção positiva.

RESULTADOS E DISCUSSÃO

Tempo computacional gasto para a execução

O programa POTION foi desenvolvido em cerca de um ano e seguiu o esquema de desenvolvimento ágil de software, no qual existe o intercâmbio constante de informações entre a equipe e a entrega semanal de resultados na forma de versões intermediárias e funcionais do software para a avaliação de desempenho. Para validarmos a paralelização de POTION, executamos o programa com um número variável de processadores utilizando o conjunto de dados de teste. A Figura 2 contém o resultado do experimento realizado.

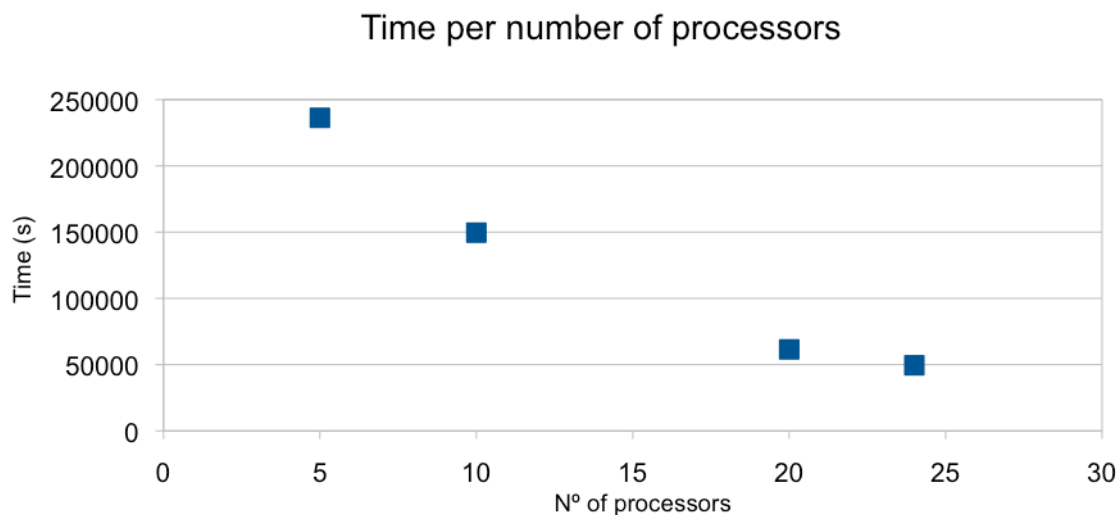


Figura 2. Tempo de computação para o conjunto de dados de teste em função do número de processadores utilizados.

É possível verificar que houve a redução aproximadamente assintótica do tempo em função do número de processadores. Através da análise do tempo computacional gasto com cada uma das etapas de execução do POTION, verificamos que o fator limitante de tempo é a execução do programa PAML que atuam como fator limitante do tempo de computação. Cerca de 80% dos grupos de homólogos encontrados como sob evidência de seleção positiva foram também detectados pelo software POTION, e no momento estamos otimizando os diversos parâmetros do mesmo para otimizarmos a detecção de grupos.

CONCLUSÃO

O software POTION, desenvolvido e validado nesse estudo, se apresentou de acordo com o que se esperava dele, uma vez que foi capaz de paralelizar a análise de grupos de genes homólogos em dois níveis (análise paralelas de grupos e de modelos estatísticos por grupo). Os resultados detectados pelo software possuem significado biológico, uma vez que diversos grupos de genes sob evidência de seleção positiva encontrados em outros estudos foram também prontamente detectados por POTION.

AGRADECIMENTOS

À Embrapa Informática Agropecuária, pela bolsa concedida e pela oportunidade de estágio.

REFERÊNCIAS

1. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol*, **26**, 1135-1145.
2. Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, **9**, 387-402.
3. Dharia, N.V., Bright, A.T., Westenberger, S.J., Barnes, S.W., Batalov, S., Kuhen, K., Borboa, R., Federe, G.C., McClean, C.M., Vinetz, J.M. *et al.* (2010) Whole-genome sequencing and microarray analysis of ex vivo Plasmodium vivax reveal selective pressure on putative drug resistance genes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 20045-20050.
4. Aguilera, G., Refregier, G., Yockteng, R., Fournier, E. and Giraud, T. (2009) Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infect Genet Evol*, **9**, 656-670.
5. Aguilera, G., Lengelle, J., Marthey, S., Chiapello, H., Rodolphe, F., Gendrault, A., Yockteng, R., Vercken, E., Devier, B., Fontaine, M.C. *et al.* (2010) Finding candidate genes under positive selection in Non-model species: examples of genes involved in host specialization in pathogens. *Mol Ecol*, **19**, 292-306.
6. Yang, Z. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, **15**, 496-503.
7. Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res*, **11**, 863-874.
8. Nielsen, R. (2005) Molecular signatures of natural selection. *Annu Rev Genet*, **39**, 197-218.
9. Van Valen, L. (1974) Molecular evolution as predicted by natural selection. *J Mol Evol*, **3**, 89-101.
10. Pan, C., Kim, J., Chen, L., Wang, Q. and Lee, C. (2007) The HIV positive selection mutation database. *Nucleic Acids Res*, **35**, D371-375.

11. Chen, S.L., Hung, C.S., Xu, J., Reigstad, C.S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R.R., Ozersky, P. *et al.* (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc Natl Acad Sci U S A*, **103**, 5977-5982.
12. Mu, J., Myers, R.A., Jiang, H., Liu, S., Ricklefs, S., Waisberg, M., Chotivanich, K., Wilairatana, P., Krudsood, S., White, N.J. *et al.* (2010) *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet*, **42**, 268-271.
13. Schwalie, P.C. and Schultz, J. (2009) Positive selection in tick saliva proteins of the Salp15 family. *J Mol Evol*, **68**, 186-191.
14. Hughes, A.L. (1994) Conserved proteins as immunogens: glutathione S-transferase of *Schistosoma*. *Parasitol Today*, **10**, 149-151.
15. Proux, E., Studer, R.A., Moretti, S. and Robinson-Rechavi, M. (2009) Selectome: a database of positive selection. *Nucleic acids research*, **37**, D404-407.
16. Nickel, G.C., Tefft, D. and Adams, M.D. (2008) Human PAML browser: a database of positive selection on human genes using phylogenetic methods. *Nucleic acids research*, **36**, D800-808.
17. Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24**, 1586-1591.
18. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**, 1792-1797.
19. Retief, J.D. (2000) Phylogenetic analysis using PHYLIP. *Methods Mol Biol*, **132**, 243-258.
20. Esteban, D.J. and Hutchinson, A.P. (2011) Genes in the terminal regions of orthopoxvirus genomes experience adaptive molecular evolution. *BMC genomics*, **12**, 261.