

Descoberta não-supervisionada de possíveis regiões polimórficas em seqüências de cDNA

Wagner Arbex, Marcos Vinícius G. Barbosa da Silva

Empresa Brasileira de Pesquisa Agropecuária, Rua Eugênio do Nascimento, 610, 36038-330, Juiz de Fora, MG

E-mail: {arbex, marcos}@cnp.gl.embrapa.br

Vítor Santos Costa

Universidade do Porto, Rua Campo Alegre, 1021/1055 - Gab. 1.45, 4169-007, Porto, Portugal

E-mail: vsc@dcc.fc.up.pt

Luiz Alfredo Vidal de Carvalho

Universidade Federal do Rio de Janeiro, Centro de Tecnologia - Bloco H-319, 21945-970, Rio de Janeiro, RJ

E-mail: alfredo@cos.ufrj.br

O volume de dados dos projetos de seqüenciamento de genomas obrigou a criação de novas ferramentas de computação com alto poder de processamento e desenvolvidas na forma de *pipelines* para automatizar o tratamento da informação genômica, passando pela aquisição, identificação, análise, prospecção, até o armazenamento da mesma. Para a obtenção de bons resultados na investigação de polimorfismos ou de polimorfismos de base única (*single nucleotide polymorphisms* - SNPs), devem ser integradas a essas ferramentas rotinas com tal finalidade e que forneçam saídas de fácil leitura. O *script* polymorp.pl é uma ferramenta de bioinformática que faz a prospecção de polimorfismos em seqüências de cDNA, a partir de sua integração com o *pipeline* phredPhrap, um *script* de leitura e montagem de seqüências genômicas, utilizado em projetos genoma de diversos organismos, que otimiza o funcionamento dos programas Phrep, phd2fasta, Cross_match e Phrap - desenvolvidos na Universidade de Washington - permitindo a utilização desses em grandes volumes de dados. A investigação de polimorfismos envolve (a) identificar o polimorfismo em uma seqüência, (b) verificar se o polimorfismo não é um erro na seqüência, (c) verificar se o polimorfismo altera a formação das proteínas e (d) verificar se a “nova” proteína formada, quando combinada com outras, manifesta ou inibe alguma característica. O polymorp.pl objetiva identificar polimorfismos em uma dada seqüência, questão (a), e fornecer medidas e estatísticas para auxiliar na verificação desses, questão (b). O phredPhrap gera as seqüências genéticas e seus atributos, que podem auxiliar em buscas supervisionadas sobre as mesmas, entretanto, o polymorp.pl trata somente essas seqüências, selecionando partes - subseqüências - consideradas de boa qualidade e identificando polimorfismos por comparação entre essas, visto que estão alinhadas. A seleção das subseqüências é feita com base na região cujo o consenso, obtido com o alinhamento das seqüências, em tese, apresenta melhor credibilidade. Como as seqüências são oriundas de cDNA, e, portanto, sofreram transcrição reversa, deve ser considerada a possibilidade de que nucleotídeos podem ter sido transcritos incorretamente, gerando falsos polimorfismos. Esses falsos polimorfismos são ainda mais críticos em SNPs e uma das formas de se identificar e eliminar tais problemas passa pela frequência alélica mínima. O polymorp.pl, em sua saída, além de representar graficamente os polimorfismos encontrados, apresenta os números absolutos e percentuais dos aminoácidos das subseqüências pesquisadas, auxiliando no cálculo da frequência alélica mínima.

Keywords: Ferramenta de busca não-supervisionada, Polimorfismo e Polimorfismo de base única