WB2016
Workshop de Bioinformática da UTFPR - 2016

Cornélio Procópio - PR
5 a 7 de outubro de 2016

ISBN: 978-85-7014-180-4
EDITORA UTFPR

DESENVOLVIMENTO DE UMA FERRAMENTA PARA IDENTIFICAR REGIÕES CODIFICADORAS NOS TRANSCRITOS DO FUNGO *Phakopsora pachyrhizi*

Cynara Leão GARCIA[1], Francismar Corrêa MARCELINO-GUIMARAES[2], André Yoshiaki KASHIWABARA[1]

[1] Programa de Pós-graduação em Bioinformática (PPGBIOINFO), Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Brasil
cynara@alunos.utfpr.edu.br
[2] Empresa Brasileira de Pesquisa Agropecuária - Embrapa, Londrina, Brasil

A soja representa uma das principais culturas agrícolas no Brasil em termos de economia, desenvolvimento e empregabilidade, sendo líder nacional em produção e área cultivada. No entanto, inúmeros fatores podem colocar em risco esse cenário, fatores esses que estão relacionados às condições climáticas e doenças. Dentre as doenças conhecidas, a mais preocupante é a ferrugem asiática da soja (FAS), pelo fungo *Phakopsora pachyrhizi*, que pode provocar perdas na produção da soja de 30% a 75%. Apesar de sua importância, o genoma completo desse fungo ainda não está disponível; assim como outras espécies de ferrugem, espera-se que o genoma *P. pachyrhizi* seja altamente complexo, com conteúdo altamente repetitivo, o que dificulta então sua montagem, além de ser estimado em cerca de 500 a 800 Mb, representando um extremo quando comparado aos genomas de outros fungos já sequenciados. Por outro lado, vários trabalhos visando identificar sequencias expressas do fungo já foram descritos, como o transcriptoma composto por 36.350 *contigs*, resultantes de montagem *ab-initio* de sequência única de *P. pachyrhizi*, obtidos de microdissecção a laser de lesão aos 10 dias de infecção. Apesar do número significativo de contigs descritos, muitos ainda não apresentam anotação e quantidade de *non hits* ainda é elevada. Dessa forma, ferramentas de bioinformática que possam auxiliar na predição de sequências codificadoras expressas se tornam extremamente úteis. Neste trabalho, dois programas mais citados pelo Google Scholar: ESTscan e OrfPredictor foram comparados, utilizando como entrada a base de dados de sequências contendo UTRs e CDSs de genes preditos que compõe o genoma da planta *Arabidopsis thaliana* obtida do banco de dados TAIR. Ambos os programas ESTScan e o OrfPredictor apresentaram uma taxa elevada de falsos positivos (41,90% e 87,85% respectivamente) de modo que faz-se necessária a implementação de um novo método para a análise de sequências transcritas. Assim, este projeto visa desenvolver um modelo de análise de transcritos utilizando o ToPS, com base na representação de regiões codificadoras utilizando Cadeias Generalizadas de Markov (GHMM). Esta abordagem amplamente utilizada em preditores de genes ainda foi pouco explorada por preditores de regiões codificadoras em sequências expressas. Adicionalmente, será também utilizado o algoritmo Viterbi para segmentar os transcritos em regiões codificadoras e regiões não traduzidas, utilizando-se de uma heurística capaz de efetuar a identificação dos erros de sequenciamento. Finalmente, o modelo deve ser capaz de ser aplicado tanto à problemática do *P. pachyrhizi* e generalizável para outras espécies.

Palavras-chave: Transcritos, ESTs, *Phakopsora pachyrhizi,* GHMM, bioinformática.

Agradecimentos: EMBRAPA