Abordagem computacional para a identificação de elementos cis-regulatórios no genoma da soja

VITORINO, J.C.[1]; SILLA, P.R.[1]; CAMARGO-BRUNETTO, M.A. de O.[1]; BINNECK, E.[2] .
[1]Universidade Estadual de Londrina – UEL, Departamento de Computação, Caixa Postal 6001, CEP 86051-990, Londrina-PR, josue.crispim@gmail.com; [2]Embrapa Soja, Caixa Postal 231, CEP 86001-970, Londrina-PR.

Introdução

O primeiro passo na expressão de um gene é a transcrição. No processo de transcrição muitos fatores internos ou externos, na célula, podem influenciar induzindo ou reprimindo a expressão dos diversos genes codificados no genoma do organismo. Fatores externos desafiantes, como estresses bióticos e abióticos, até mecanismos moleculares intrínsecos podem desencadear, direta ou indiretamente, a ativação da expressão gênica espaço-temporal. A região promotora e seus elementos cis-regulatórios, presentes na estrutura de cada gene, são fundamentais para o processo de transcrição. Por isso, entre outros aspectos, o conhecimento dos elementos cis-regulatórios é essencial para o entendimento da regulação de um determinado gene (Wang et al, 2009) e esse conhecimento é fundamental para interpretar e modelar as respostas de uma célula a diversos estímulos (Wasserman et al, 2004).

Os elementos cis, geralmente são pequenos segmentos de DNA (5 a 20 nucleotídeos) encontrados na região promotora, que fica upstream do sitio de inicio da transcrição dos genes que regulam. Devido ao seu pequeno tamanho, a identificação de um elemento cis em um gene é uma tarefa difícil. A Figura 1 mostra a região promotora de um gene com sítios de ligação de fatores de transcrição, destacados em roxo, upstream do sitio de início da transcrição.

Na região promotora, determinadas proteínas, conhecidas como fatores de transcrição, reconhecem e ligam-se aos elementos cis, formando um complexo que interfere no posicionamento correto da RNA-polimerase II no promotor, na separação das fitas de DNA para permitir o início da transcrição, e liberam a RNA-polimerase II do promotor quando a transcrição se inicia.

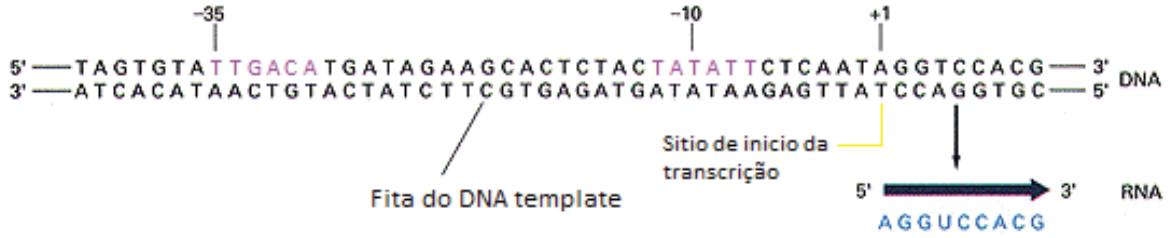

Figura 1. Região promotora, com dois elementos regulatórios.

A identificação experimental de elementos *cis* é cara, demorada e difícil. Isso faz dos métodos computacionais as ferramentas ideais para predizer elementos *cis*, antecipando os estudos experimentais de regulação da expressão gênica.

Existem vários algoritmos computacionais desenvolvidos para encontrar elementos *cis* nos genes de diversos organismos, os quais podem ser divididos em dois grupos: (1) os algoritmos baseados em sequências promotoras de genes co-regulados e (2) algoritmos baseados em rastros filogenéticos (Das et al, 2007).

Os algoritmos baseados em genes co-regulados ainda podem ser divididos em dois subgrupos: de predição baseada em palavras e predição probabilística.

Algoritmos de predição baseada em palavras computam todas as possíveis subsequências que podem ocorrer, através de diferentes sequências promotoras. Encontrado o número de frequência de uma subsequência, este deve ser comparado com o número de frequência esperada. Depois, são utilizados métodos estatísticos para avaliar a significância da sequência observada (Rombauts et al, 2003).

Os algoritmos de predição probabilística, para encontrar os elementos *cis*, constroem um alinhamento múltiplo alinhando localmente pequenas regiões conservadas. Os métodos probabilísticos começam com uma subsequência modelo representada através de uma matriz de peso que é alterada em séries de iterações até encontrar uma pontuação ótima (Rombauts et al, 2003).

Os algoritmos baseados em rastros filogenéticos assumem que elementos cis são regiões conservadas no DNA e não sofreram muitas mutações ao longo da evolução. Esses algoritmos comparam sequências promotoras de genes ortólogos de múltiplas espécies para identificar os elementos cis.

Atualmente existem poucas ferramentas dedicadas à descoberta de elementos cis em plantas, a maior parte das soluções são baseadas em fungos, mamíferos e insetos como a *Drosophila*. Das ferramentas dedicadas a plantas a maior parte é baseada na planta modelo Arabidopsis.

Objetivos

Os objetivos deste trabalho são, nesta ordem:

- Criar um método para encontrar as sequências promotoras do gene, onde se deseja encontrar os elementos *cis*.
- Programar um algoritmo utilizando a teoria dos rastros filogenéticos, para a predição de elementos *cis*.
- Automatizar o método de busca das sequências com *scripts*.
- Construir um banco de dados com os elementos *cis* preditos.

Metodologia

Para o desenvolvimento deste trabalho serão realizados estudos de algoritmos baseados em rastros filogenéticos, como o Footprinter (Blanchette et al, 2002); algoritmos projetados para encontrar elementos *cis* mais parcimoniosos dinamicamente de tamanho *k*, como o Footer (Benos et al, 2007); assim como algoritmos baseados tanto em rastros filogenético como em genes co-regulados, como os algoritmos PhyME (Sinha et al. 2004) e PhyloCon (Wang et al. 2003).

A segunda etapa será a de desenvolver uma ferramenta baseada nos estudos realizados. Serão criadas técnicas para encontrar e extrair as sequências promotoras dos genes da soja, e dos genes ortólogos de espécies próximas à soja, utilizando ferramentas como GBrowse e BioMart. Essas sequências serão usadas no processo de predição computacional de elementos *cis* no genoma da soja.

Havendo a possibilidade de tempo, será automatizado o processo de busca de sequências promotoras, com o desenvolvimento de *scripts* para acessar diretamente as ferramentas mencionadas acima, através de uma interface Web.

Para armazenar os elementos *cis*, será criado um banco de dados em MySQL. Esse banco de dados poderá ser acessado via Web, para que os elementos *cis* preditos possam ser avaliados e validados em análises laboratoriais.

Referências

DAS M.K.; DAI H.K. A survey of DNA motif finding algorithms. **BMC Bioinformatics**, v. 8, p. 1471-2105, 2007.

BENOS P. V.; CORCORAN D. L.; FEINGOLD E. Web-based identification of evolutionary conserved DNA cis-regulatory elements. **Methods in Molecular Biology**, v. 395, p. 425-436, 2007.

BLANCHETTE M.; SCHWIKOWSKI B.; TOMPA M. Algorithms for phylogenetic footprinting. **Journal of Computational Biology**, v. 9, n. 2, p. 211-223, 2002.

ROMBAUTS S.; FLORQUIN K.; LESCOT M.; MARCHAL K.; ROUZÉ P.; PEER V. Y. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. **Plant Physiology**, v. 132, n. 3, p. 1162-1176, 2003.

SINHA S.; BLANCHETTE M.; TOMPA M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. **BMC Bioinformatics**, v. 5, n. 1, p. 170, 2004.

WANG T.; STORMO G.D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. **Bioinformatics**, v. 19, n. 18, p. 2369-2380, 2003.

WANG, X.; HABERER, G.; MAYER, K. Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. **BMC Genomics**, v. 10, p. 1471-2164, 2009.

WASSERMAN W.W.; SANDELIN A. Applied bioinformatics for the identification of regulatory elements. **Natural Reviews Genetics**, v. 5, p. 276-287, 2004.