PROJETO DE TCC

Título:
Aluno: Josué Crispim Vitorino
Professora Orientadora: Maria Angélica de Oliveira Camargo Brunetto

Sumário

1 RESUMO p. 3

2 FUNDAMENTAÇÃO TEÓRICO-METODOLÓGICA p. 4

3 JUSTIFICATIVA CIRCUNSTANCIADA p. 8

Referências Bibliográficas p. 9

1 RESUMO

Nos últimos anos, alguns fatores de transcrição que regulam a expressão de vários genes relacionados com o estresse foram descobertos. Esses fatores de transcrição são subdivididos em várias classes como a Dehydration Responsive Element Binding Proteins (DREB), que está relacionada com a seca e a desidratação da planta. O entendimento dos DREBs é importante para o desenvolvimento de plantas com tolerância a estresses abióticos como a seca, alta salinidade e baixa temperatura. Esse projeto apresenta abordagens computacionais que serão desenvolvidas no empenho de encontrar os elementos regulatórios na soja, que são ativados pelos fatores de transcrição pertencentes classe DREB.

2 FUNDAMENTAÇÃO TEÓRICO-METODOLÓGICA

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 transcrição consiste na formação do RNA a partir do DNA, para que a transcrição ocorra é necessário a ação de uma enzima chamada RNA-polimerase, essa enzima se conecta na sequência de DNA, próximo a região onde está o local de inicio da transcrição (LIT) e se move sobre o DNA, no sentido contrario ao LIT formando o RNA, a transcrição é iniciada exatamente após o LIT. Por sua vez para que a RNA-polimerase consiga se conectar no DNA é imprescindível a ação conjunta de proteínas especiais que se conectam a determinados segmentos de DNA.

Essas proteínas são chamadas de fatores de transcrição (TFs), elas podem se conectar distante da RNA-polimerase ou próximo a RNA-polimerase formando um complexo de vários fatores de transcrição juntamente com a RNA-polimerase. Os segmentos de DNA em que os TFs se ligam são pequenos(5 a 20 nucleotídeos), e são chamados de elementos regulatórios. Geralmente os elementos regulatórios estão localizados em uma região antes do local de inicio da transcrição, como mostrado na figura 1. Esta região recebe o nome de região promotora, ela será responsável de promover um gene, para ser expresso. A ligação dos TFs nos elementos regulatórios interfere no posicionamento correto da RNA-polimerase, na separação das fitas de DNA para permitir o inicio da transcrição, e na liberação da RNA-polimerase quando a transcrição se inicia e consequentemente na transcrição de um gene. Quando ocorre a transcrição parte do RNA transcrito ira formar posteriormente proteínas, que são essenciais para a sobrevivência do organismo, estas proteínas podem até mesmo ser TFs que irão ativar outros genes.

A transcrição esta diretamente ligada a respostas das células a estímulos, como mudanças hormonais internamente em um organismo, ou externamente como estresses abióticos e bióticos. Os estresses abióticos são causados por fatores não vivos como a alteração de temperatura e mudança climática. Os estresses bióticos são causados por organismos vivos como bactérias, vírus, parasitas e insetos.

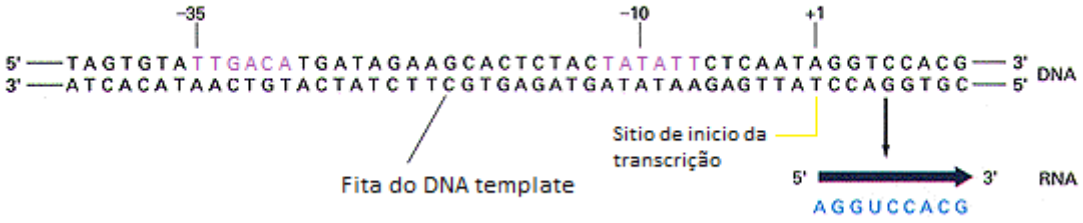

Figura 1. Região promotora, com dois elementos regulatórios.

No amplo conjunto de fatores de transcrição, existem aqueles que quando ligados nos elementos regulatórios irão ativar as respostas da célula a estresses abióticos. Em organismos vegetais os estresses abióticos que mais prejudicam são: a seca, alta salinização e baixas temperaturas. Esses fatores de transcrição quando ligados aos elementos regulatórios irão desencadear uma série de eventos que resultara na proteção da célula e sua tolerância a estresses. Na *Arabdopsis*, uma planta modelo amplamente utilizada em pesquisas de genética molecular nas plantas, os fatores de transcrição relacionados a estresses abióticos são agrupados em classes (ou famílias), uma das principais classes é a Dehydration Responsive Element Binding Proteins (DREB), que por sua vez pertence a família Ethylene Responsive Element (ERF), uma importante família de fatores de transcrição de respostas a estresses. O DREB é subdividido em duas subclasses: DREB1/CBF e DREB2 que são induzidas pelo frio e desidratação, respectivamente (AGARWAL et al.,).

Segundo Agarwal et al.(AGARWAL et al.,) estresses abióticos e bióticos influenciam negativamente na sobrevivência e na larga produção de grãos. Culturas como soja, arroz e trigo que são amplamente usadas na alimentação mundial são prejudicadas pelos estresses que muitas vezes impedem uma alta produtividade. O entendimento dos DREBs na regulação de um gene é de grande importância para o desenvolvimento de plantas tolerantes a estresses.

Existem vários métodos computacionais desenvolvidos para encontrar elementos regulatórios nos genes de diversos organismos, Das e Dai (DAS; DAI, 2007) classificou os métodos existentes em três grupos:

- Os baseados em sequências promotoras de genes que são regulados pelos mesmos fatores

de transcrição (genes co-regulados), estes métodos se concentram em apenas um único genoma.

- Os que utilizam sequencias promotoras ortólogas, que são sequencias de DNA similares a várias espécies, indicando que estas espécies derivaram de um ancestral comum, também chamados de métodos de rastros filogenéticos.
- Os métodos que combinam rastros filogenéticos e sequencias promotoras de genes co-regulados.

Os métodos 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 modelos de predição probabilística, geralmente utilizam matriz de peso e os parâmetros do modelo são estimado usando o princípio de inferência bayesiana. Há varias implementações baseadas no método probabilístico, entre estas técnicas destacam as técnicas estatísticas como *EM* método e *Gibb sampling*, técnicas de aprendizado de máquina e técnicas de *Ensemble* (DAS; DAI, 2007).

Os algoritmos baseados em rastros filogenéticos assumem que elementos regulatórios 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 regulatórios.

Por último os algoritmos que combinam as técnicas probabilísticas e de rastros filogenéticos, que integram dois importantes aspectos dos elementos regulatórios, a sobre-representação e a conservação dos elementos regulatórios entre múltiplas espécies (DAS; DAI, 2007). Algumas das implementações dos modelos citados estão listadas na tabela 2.1.

Dos modelos apresentados, a predição baseada em palavras mostrou-se eficiente na busca de elementos regulatórios em organismos eucarióticos mas problemática em organismos procarióticos, devido ao tamanho dos elementos regulatórios nos eucarióticos serem menores do que nos procarióticos. A predição baseada em rastros filogenéticos, mostrou-se muito eficiente em organismos procarióticos, mas é necessário as sequencias de varias espécies, para

fazer as comparações, com o avanço do sequenciamento do genoma das espécies este método ficara cada vez mais expressivo. O modelo de predição probabilística também tem-se mostrado eficiente na busca de elementos regulatórios em grandes genomas, uma das técnicas que se destaca com grande eficiência na busca de elementos regulatórios dentro dos modelos que utilizam aprendizado de máquina é a de *support vector machine* (SVM).

Até o presente momento poucos foram os trabalhos que se dedicaram exclusivamente na busca DREBs nas plantas. Recentemente Wang et al. (WANG et al., 2009) criou um método utilizando SVM para encontrar genes que eram expressos quando expostos a fatores abióticos na *Arabidopsis*. Eles utilizaram sequencias de DNA de regiões promotoras de genes da *Arabidopsis*, que possuía elementos regulatórios que se conectavam a DREBs como dados positivos e sequencias aleatórias da região promotora representando os dados negativos, primeiramente foi aplicado o algoritmo HexDiff (CHAN; KIBLER, 2005) para análise dos hexâmeros sobre-representados nas sequencias e então as sequencias promotoras dos genes foram classificadas com o SVM, discriminando os as sequencias de genes que eram alvos de DREBs das que não eram.

Algoritmos	Referências
Algoritmos de predição baseada em palavras	
Oligo-Analysis	(HELDEN; ANDRé; COLLADO-VIDES, 1998)
YMF	(SINHA; TOMPA, 2003)
MITRA	(ESKIN; PEVZNER, 2002)
Algoritmos de predição probabilística	
MEME	(BAILEY et al., 2006)
Gibbs sampling	(LAWRENCE et al., 1993)
AlignACE	(ROTH et al., 1998)
Motif Sampler	(THIJS et al., 2002)
Algoritmos baseados em rastros filogenéticos	
Footprinter	(BLANCHETTE; TOMPA, 2002)
PHYLONET	(WANG, 2005)
PhyloScan	(CARMACK et al., 2007)
Algoritmos baseados em rastros filogenéticos e predição probabilística	
OrthoMEME	(CARMACK et al., 2007)
PhyloCon	(WANG; STORMO, 2003)
PhyME	(SINHA; BLANCHETTE; TOMPA, 2004)

Tabela 2.1: Implementações de modelos de predição

Atualmente existem poucas ferramentas dedicadas à descoberta de elementos regulatórios 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.

3 JUSTIFICATIVA CIRCUNSTANCIADA

A região promotora e seus elementos regulatórios, presentes na estrutura de cada gene, são fundamentais para o processo de transcrição de um gene. Por isso, entre outros aspectos, o conhecimento dos elementos regulatórios e dos fatores de transcrição é essencial para o entendimento da regulação de um determinado gene (**??**) e um passo fundamental na construo da rede de regulação de um gene. Esse conhecimento fundamental para interpretar e modelar as respostas de uma célula a diversos estímulos (**??**).

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

Referências Bibliográficas

AGARWAL, P. K. et al. Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. *Plant Cell Reports*, v. 25, n. 12, p. 1263–1274. ISSN 0721-7714. Disponível em: <http://dx.doi.org/10.1007/s12190-008-0204-7>.

BAILEY, T. L. et al. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, Institute of Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia. t.bailey@imb.uq.edu.au, v. 34, n. suppl 2, p. W369–W373, jul. 2006. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkl198>.

BLANCHETTE, M.; TOMPA, M. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Research*, v. 12, n. 5, p. 739–748, maio 2002. ISSN 1088-9051. Disponível em: <http://dx.doi.org/10.1101/gr.6902>.

CARMACK, C. S. et al. PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms for Molecular Biology*, v. 2, n. 1, p. 1+, jan. 2007. ISSN 1748-7188. Disponível em: <http://dx.doi.org/10.1186/1748-7188-2-1>.

CHAN, B.; KIBLER, D. Using hexamers to predict cis-regulatory motifs in drosophila. *BMC Bioinformatics*, v. 6, n. 1, p. 262+, out. 2005. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186/1471-2105-6-262>.

DAS, M.; DAI, H. K. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, v. 8, n. Suppl 7, p. S21+, 2007. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186-/1471-2105-8-S7-S21>.

ESKIN, E.; PEVZNER, P. A. Finding composite regulatory patterns in DNA sequences. *Bioinformatics (Oxford, England)*, Department of Computer Science, Columbia University, New York, 10027 NY, USA. eeskin@cs.columbia.edu, v. 18 Suppl 1, 2002. ISSN 1367-4803. Disponível em: <http://view.ncbi.nlm.nih.gov/pubmed/12169566>.

HELDEN, J. van; ANDRé, B.; COLLADO-VIDES, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies1. *Journal of Molecular Biology*, Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, AP565A Cuernavaca, Morelos, 62100, México. jvanheld@ebi.ac.uk, v. 281, n. 5, p. 827–842, set. 1998. ISSN 00222836. Disponível em: <http://dx.doi.org/10.1006/jmbi.1998.1947>.

LAWRENCE, C. E. et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science (New York, N.Y.)*, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894., v. 262, n. 5131, p. 208–214, out. 1993. ISSN 0036-8075. Disponível em: <http://dx.doi.org/10.1126/science.8211139>.

ROMBAUTS, S. et al. Computational Approaches to Identify Promoters and cis-Regulatory Elements in Plant Genomes. *Plant Physiol.*, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, B-9000 Gent, Belgium., v. 132, n. 3, p. 1162–1176, jul. 2003. ISSN 0032-0889. Disponível em: <http://dx.doi.org/10.1104/pp.102.017715>.

ROTH, F. P. et al. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, Nature Publishing Group, Harvard University Graduate Biophysics Program and Harvard Medical School Department of Genetics, Boston, MA 02115, USA., v. 16, n. 10, p. 939–945, out. 1998. ISSN 1087-0156. Disponível em: <http://dx.doi.org/10.1038/nbt1098-939>.

SINHA, S.; BLANCHETTE, M.; TOMPA, M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10021, USA. saurabh@lonnrot.rockefeller.edu, v. 5, n. 1, p. 170+, out. 2004. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186/1471-2105-5-170>.

SINHA, S.; TOMPA, M. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, v. 31, n. 13, p. 3586–3588, jul. 2003. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkg618>.

THIJS, G. et al. A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes. *Journal of Computational Biology*, ESAT-SCD, KULeuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. GertThijs@esat.kuleuven.ac.be, v. 9, n. 2, p. 447–464, abr. 2002. ISSN 1066-5277. Disponível em: <http://dx.doi.org/10-.1089/10665270252935566>.

WANG, S. et al. An <i>in silico</i> strategy identified the target gene candidates regulated by dehydration responsive element binding proteins (DREBs) in <i>Arabidopsis</i> genome. *Plant Molecular Biology*, Springer Netherlands, v. 69, n. 1, p. 167–178, jan. 2009. ISSN 0167-4412. Disponível em: <http://dx.doi.org/10-.1007/s11103-008-9414-5>.

WANG, T. Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proceedings of the National Academy of Sciences*, v. 102, n. 48, p. 17400–17405, nov. 2005. ISSN 0027-8424. Disponível em: <http://dx.doi.org/10.1073/pnas.0505147102>.

WANG, T.; STORMO, G. D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, Department of Genetics, Washington University Medical School, St. Louis, MO 63110, USA., v. 19, n. 18, p. 2369–2380, dez. 2003. ISSN 1367-4803. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btg329>.