PROJETO DE TCC

Título: Abordagem computacional para a identificação de elementos regulatórios na soja.
Aluno: Josué Crispim Vitorino
Professora Orientadora: Maria Angélica de Oliveira Camargo Brunetto

Sumário

1 RESUMO

Nos últimos anos muitos estudos, estão sendo realizados para entender a regulação de um gene. O completo entendimento desta será um grande avanço na genética. Em particular nas plantas, este entendimento poderá ajudar, em pesquisas de melhoramentos genéticos.

Para entender a regulação de um gene, que é um complexo processo que envolve diversos fatores dentro das células, estudos estão sendo feitos para a identificação de sequências específicas no DNA, chamadas de elementos regulatórios. Estes elementos, juntamente com os fatores de transcrição, são responsáveis pelo início da transcrição em uma célula. Eles funcionam como mecanismos de resposta das células a eventos interna e externamente nas células, como: mudanças hormonais, elevação da temperatura e seca. Estas mudanças muitas vezes afetam negativamente as plantas, interferindo na produtividade da mesma. Com o avanço nas descobertas dos elementos regulatórios, poderão ser desenvolvidas plantas mais resistentes a essas condições adversas. Esse projeto apresenta abordagens computacionais que serão desenvolvidas no empenho de encontrar os elementos regulatórios, que são ativados quando a planta é exposta a estresses. Os estudos serão conduzidos no genoma da soja, uma cultura amplamente utilizada e importante para a economia nacional.

2 FUNDAMENTAÇÃO TEÓRICO-METODOLÓGICA

Atualmente com o sequenciamento de genomas de varias espécies, muitos estudos estão sendo realizados para descifrar o funcionamento e comportamento das células. Muitos avanços no estudo envolvendo genomas foram alcançados, como por exemplo a caracterização de genes. Mas ainda resta muito trabalho para entender a informação inserida no genoma de um organismo.

Uma das atividades da célula, que ainda está distante de ser completamente entendida, é a regulação da expressão gênica. O entendimento da regulação dos genes irá trazer benefícios para as pesquisas como a de tratamento e prevenção de doenças. Para entendê-la é essencial a identificação dos elementos regulatórios, uma vez que para que a transcrição de um gene ocorra é necessário a ação destes elementos. O entendimento dos elementos regulatórios é um passo fundamental para a compreensão da regulação de um gene.

Para a identificação dos elementos regulatórios, foram propostos vários métodos, mas ainda não existe um método livre de falhas. Nas seções seguintes serão explicados o início da regulação de um gene e o papel dos elementos regulatórios na regulação de um gene. Também serão apresentados as metodologias computacionais que são adotadas atualmente para a identificação dos elementos regulatórios.

2.1 Início da regulação de um gene

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.

2.1.1 Ácidos nucleicos

Os ácidos nucleicos são moléculas que exercem um importante papel nos organismos vivos. Neles estão contidos as informações genéticas da célula. É a partir deles que as células recebem instruções de quais proteínas sintetizar e em que quantidade. Essa informação é decifrada através do código genético, cuja a tradução resulta na síntese de proteínas (ZAHA, 2000). Existem dois tipos de ácidos nucleicos: ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA), ambos os ácidos são compostos por nucleotídeos. Os nucleotídeos são formados a partir de três componentes químicos: um açúcar, uma base nitrogenada e um ácido fosfórico (figura 1). Os nucleotídeos estão ligados entre eles formando uma sequência linear (figura 2).

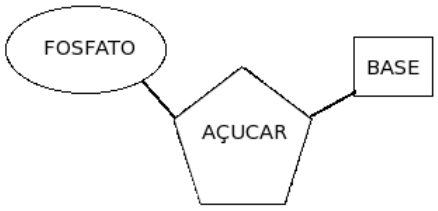

Figura 1: Componentes de um nucleotídeo

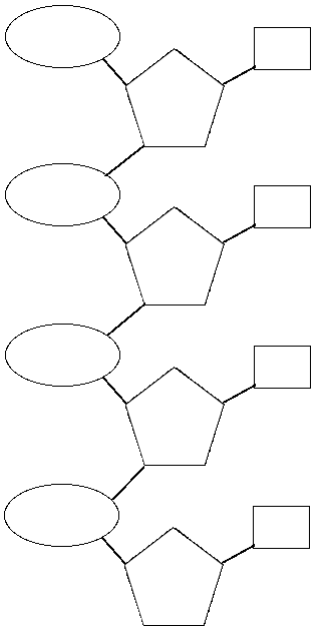

Figura 2: Sequência linear de nucleotídeos ligados

Existem duas importantes diferenças entre o RNA e o DNA, que são: o tipo de açúcar e as bases nitrogenadas. O açúcar no DNA é a desoxirribose e no RNA a ribose (figura 3). Em cada ácido nucleico são encontradas quatros bases nitrogenadas, sendo que três delas são compartilhadas entre o RNA e DNA são elas: adenina (A), guanina (G) e citosina (C). A base

timina (T), é encontrada só no DNA, e a uracila (U), é encontrada só no RNA. Porém existe outra grande diferença entre o RNA e o DNA no nível estrutural. O RNA geralmente existe como uma única sequência de nucleotídeos, enquanto o DNA existe como duas sequências de nucleotídeos pareadas que formam um helicoide, conhecido como dupla hélice. Na figura 4, podemos observar a estruturas dos dois ácidos. As bases nitrogenadas no DNA estão no interior da hélice, ligadas por pontes de hidrogênio formando pares de bases nitrogenadas (pb). Os únicos pares possíveis no DNA são: A ligado com T e C ligado com G (ZAHA, 2000).

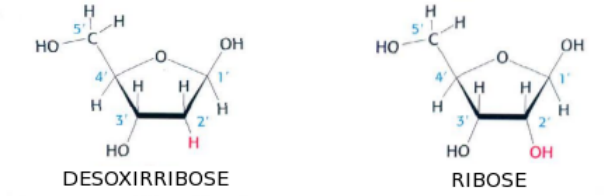

Figura 3: Tipos de açúcar encontrados nos ácidos nucleicos. (BERG; TYMOCZKO; STRYER, 2007, Adaptada))

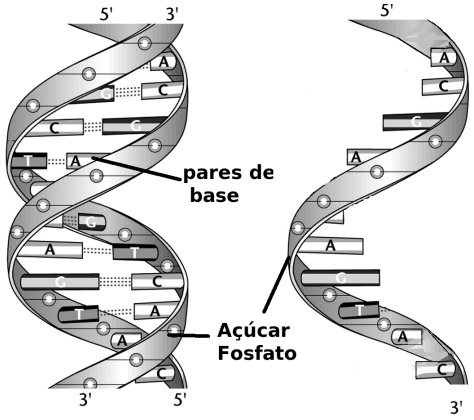

Figura 4: Estrutura do DNA e RNA. (HIGGS; ATTWOOD, 2005, Adaptada)

O RNA apesar de ter apenas uma cadeia de nucleotídeos também tem bases complementares. Na síntese de RNA, descrita com mais detalhes na seção 2.1.2, as bases que compõem a sequência do RNA são o complemento das bases copiadas do filamento (sequência de nucleotídeos) do DNA modelo, com a substituição de T por U no RNA . As bases complementares no RNA são: A ligado com U e C ligado com G.

A figura 5, mostra uma sequência de DNA e uma de RNA. Elas são comumente representas como uma palavra formada pelo alfabeto (A, G, C, T), para representações de DNA e (A, G, C, U), para representações do RNA. A leitura é feita da esquerda para a direita, no sentido indicado como 5'-> 3'. Este tipo de representação torna fácil a visualização, assim como a manipulação a nível computacional, sendo amplamente utilizado em métodos *in silico* que envolvem o DNA e o RNA.

Figura 5: Reapresentação do RNA e DNA

A interação entre o RNA e DNA ocorre quando é necessário a expressão de um gene. A expressão genética inicia-se por um processo chamado transcrição detalhado na seção 2.1.2. Neste processo ocorre a formação do RNA (síntese de RNA), a partir de um dos filamentos do DNA. A sequência de RNA formada é uma cópia exata de uma região do RNA. Esta região que é copiada é chamada de gene. Os genes são seguimentos de DNA podendo ter milhares de pares de bases. São eles que irão especificar o tipo de proteína a ser sintetizado. O processo da síntese de proteínas é conhecido como tradução. A figura 6, apresenta cada passo deste conjunto de processos, que também é comumente conhecido como o dogma central.

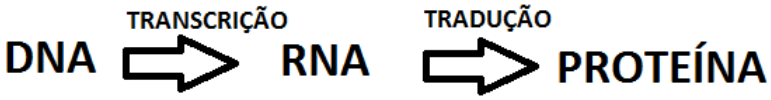

Figura 6: Principais passos da expressão de genética

2.1.2 Transcrição e a síntese de RNA

Todo o RNA é formado a partir do DNA em um processo chamado transcrição. A transcrição começa com a abertura da hélice dupla do DNA e um dos filamentos do DNA servirá como modelo para a síntese de RNA. A sequência de nucleotídeos na cadeia de RNA é determinada pelo complemento do molde do filamento de DNA (figura 7). Para que ocorra a formação do RNA é necessário a ação de uma enzima que realiza a transcrição, ela é chamada de RNA polimerase. A RNA polimerase se conecta no DNA, então ela move-se ao longo do DNA na direção 5' para a 3', formando a cadeia de RNA que vai se alongando de um nucleotídeo por vez, terminando com uma sequência de nucleotídeos exatamente complementar ao filamento de DNA usado como modelo (figura 8). Após terminada a cópia a cadeia de RNA, juntamente com a RNA polimerase, se desconectam e ocorre o fechamento da hélice do DNA. O RNA formado possuí apenas um filamento e é menor do que uma molécula de

DNA. Uma molécula de DNA no cromossomo humana pode ser maior que 250 milhões de pares de nucleotídeos, a maioria das moléculas de RNA não passam de poucos milhares de nucleotídeos.

Figura 7: RNA formado, complementar ao filamento modelo

Nos organismos eucarióticos existem três RNA polimerase, chamadas de RNA polimerase I, RNA polimerase II, RNA polimerase III. As três enzimas são similares umas com as outras. A maior diferença entre elas é o tipo de gene que elas transcrevem. RNA polimerase I e III transcrevem os genes que codificam o RNA transportador (transporta aminoácidos que se interagem com o RNA mensageiro formando as proteínas) e RNA ribossômico (um conjunto de RNA ribossômico forma o ribossomo, uma estrutura necessária na síntese da proteína), respectivamente. A RNA polimerase II transcreve a maioria dos genes, ela gera o RNA mensageiro, utilizado na formação das proteínas. Para que a RNA polimerase inicie a transcrição é necessário um grande conjunto de proteínas chamadas de fatores de transcrição gerais.

Os fatores de transcrição gerais se conectam em pontos de ligação na sequência de DNA chamados de promotores. Os promotores são pequenos segmentos de DNA, reconhecidos pela RNA polimerase, que servem como sinalizadores para o posicionamento correto da RNA polimerase. Eles estão localizados pouco antes do gene que será transcrito pela RNA polimerase. É nos promotores que a RNA polimerase se conecta, juntamente com os fatores de transcrição gerais para iniciar a transcrição. Logo após os promotores, está o local de inicio da transcrição (TSS, do inglês *transcription start site*), este é o local onde inicia-se a transcrição, ou seja o primeiro nucleotídeo a ser transcrito e a sua posição é marcada como $+1$, todos os nucleotídeos antes são marcados com suas posições com números iniciando-se pelo -1 negativos (-1, -2, -3...), e após com números positivos. Na figura 9, podemos visualizar os promotores antes do TSS, assim como a RNA polimerase e os fatores de transcrição gerais se conectando na sequência de DNA.

Os promotores seguem um padrão, e geralmente ele é o mesmo para os promotores da maioria dos genes. Por exemplo em muitas bactérias, na região promotora (região do DNA

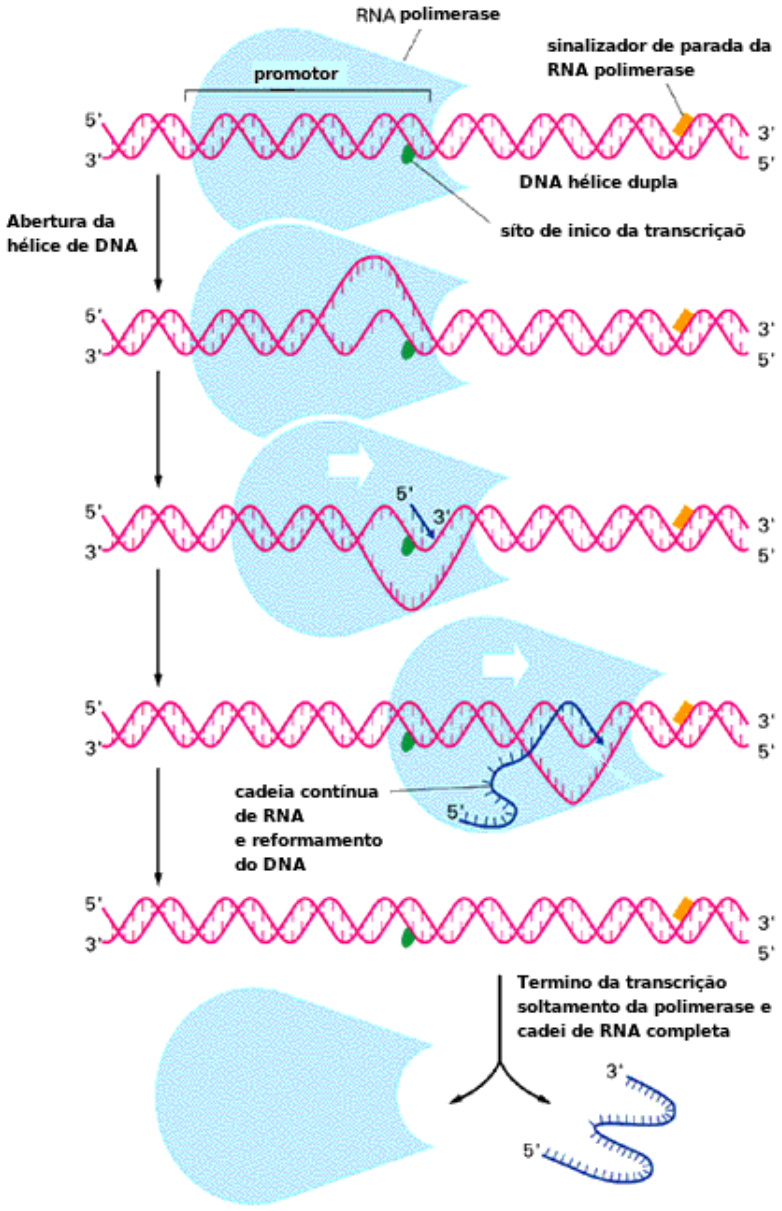

Figura 8: Formação do RNA através da RNA polimerase (HIGGS; ATTWOOD, 2005, Adaptada)

onde encontra-se os promotores de um gene) da maioria dos genes, existe uma sequência que tem como consenso a sequência **TATAATT** localizada aproximadamente na região -10. Outra sequência é a **TTGACA** localizada aproximadamente na região -35 (figura 10).

A ligação dos fatores de transcrição gerais nos promotores, ajuda no correto posicionamento da RNA polimerase, na separação dos dois filamentos de DNA para permitir o inicio da transcrição, e lançar a RNA polimerase do promotor para iniciar a transcrição. Os fatores de transcrição são ditos gerais porque, eles se ligam a todas as regiões promotoras usada pela RNA polimerase II. Estes fatores de transcrição classificados como TFII (para fatores de transcrição da polimerase II), e listados com TFIIA, TFIIB etc.

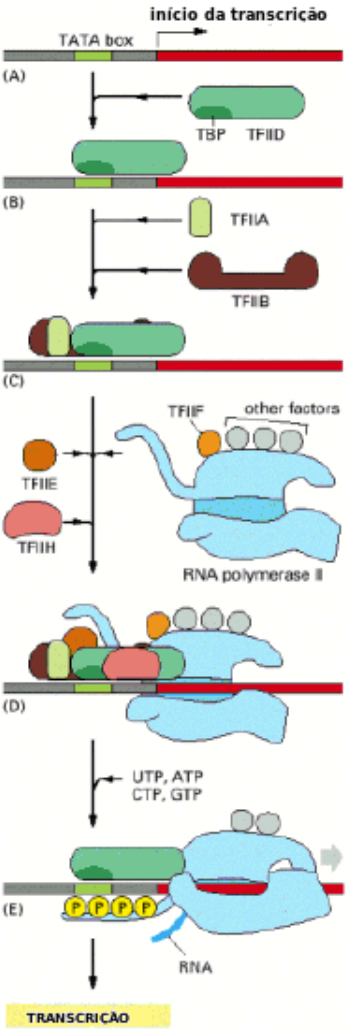

Figura 9: RNA polimerase e os fatores de transcrição gerais (ALBERTS, 2002, Adaptada)

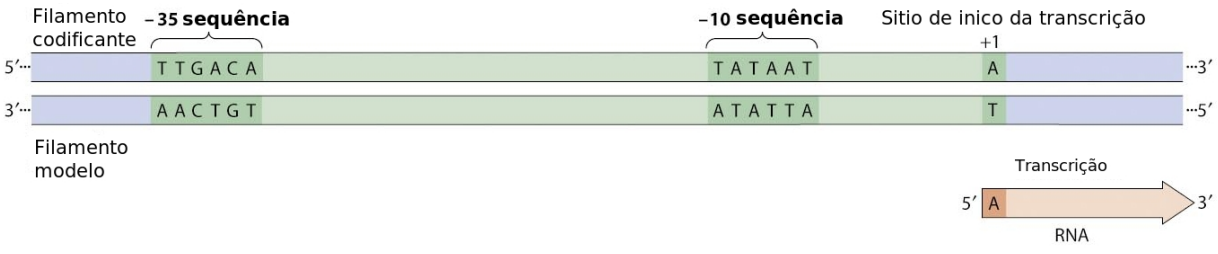

Figura 10: Região promotora e as sequências consenso

O desligamento da RNA polimerase do DNA, no fim da transcrição, não é aleatório. A RNA polimerase encontra novamente sequências consenso conhecidas como finalizadores, então ela se desconecta do DNA, que se fechará, voltando a estrutura original da dupla hélice. O RNA transcrito está pronto ou como o tRNA e rRNA ou como mensagem (mRNA) para ser traduzida em proteínas por meio da tradução.

2.1.3 Elementos regulatórios

Além dos promotores, existem outras sequências no DNA em que fatores de transcrição se conectam. Essas sequências são chamadas de elementos regulatórios(ou promotores proximais), eles ficam aproximadamente uma distância de -50 pares de base do local de início da transcrição (ZAHA, 2000). Eles se encontram assim como os promotores na região promotora de um gene e também são pequenos (5 a 20 nucleotídeos). Mas diferentes dos promotores os fatores de transcrição(TFs) que conectam a eles não são fatores de transcrição gerais, mas sim específicos. Portanto um fator de transcrição específico utilizado na expressão de um determinado gene, pode não ser o mesmo para a expressão de outro gene. Por exemplo o conjunto de fatores de transcrição de uma célula óssea no organismo humano, pode ter fatores diferentes do conjunto de TFs de uma célula do figado, uma vez que essas diferentes células podem precisar de diferentes proteínas. Essa especificidade torna os elementos regulatórios em sequências que não são consenso, em consequência os elementos regulatórios não seguem um padrão.

Existem elementos regulatórios que são ativados em resposta a estímulos como mudanças hormonais internamente em um organismo, ou externamente como: estresses abióticos que são causados por fatores não vivos como a alteração de temperatura e mudança climática, ou estresses bióticos causados por organismos vivos como bactérias, vírus, parasitas e insetos. Com a ativação do elemento regulatório ocorrera a expressão do gene, que o elemento regula, o gene será transcrito no RNA que posteriormente será traduzido, gerando proteínas para suprir as necessidades do organismo.

2.1.4 Fatores de transcrição que ativam elementos regulatórios em resposta a estresses abióticos

No amplo conjunto de fatores de transcrição, existem aqueles que quando ligados nos elementos regulatórios irão ativar as respostas da célula a estresses abióticos. O estress abiótico afeta diversos organismos, mas em especial os organismos vegetais que são dependentes de fatores ambientais, são os mais afetados. Em organismos vegetais os estresses abióticos que mais prejudicam são: a seca, alta salinização e baixas temperaturas.

A ligação entre os fatores de transcrição que estão relacionados a expressão de um gene (ou conjunto de genes), com elementos regulatórios, no momento em que o organismo é exposto a um estresse, a célula produzira proteínas, que irão agir na proteção da célula.

Na Arabdopsis, uma planta modelo amplamente utilizada em pesquisas de genética

molecular nas plantas. Com estudos em laboratório foram encontrados alguns fatores de transcrição relacionados a estresses abióticos, eles são agrupados em classes (ou famílias). Atualmente uma das classes que vem sendo objeto de muitos estudos, devido a sua resposta a diversos estresses, 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. Os DREBs desempenham um importante papel na proteção de algumas plantas, provendo tolerância a estresses e respondendo a diferentes condições de estresses, como: frio, alta salinidade e seca (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.

2.2 Métodos computacionais para a predição de elementos regulatórios

Atualmente vários métodos computacionais foram desenvolvidos para detectar elementos regulatórios. Apesar dos avanços para encontrar elementos relatórios, a busca *in silico* não é tão precisa, diferente por exemplo da classificação de genes, gerando muitos resultados falsos. Isto deixa em aberto um vasto campo para ser explorado (ROMBAUTS et al., 2003). Um dos principais problemas na predição de elementos regulatórios, é que eles são definidos funcionalmente e não estruturalmente, limitando os meios para modelá-los (ROMBAUTS et al., 2003). 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.

Dos métodos de predição existentes, Das e Dai (DAS; DAI, 2007) os classificaram 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 sequências promotoras de genes ortólogos, que são sequências de DNA similares a várias espécies, indicando que estas espécies derivaram de um ancestral co-

mum, também chamados de métodos de rastros filogenéticos.

- Os métodos que combinam rastros filogenéticos e sequências 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.

2.2.1 Algoritmos de predição baseada em palavras

Os algoritmos de predição baseada em palavras, geralmente utilizam de enumeração exaustiva, computando todas as possíveis subsequências que podem ocorrer, através de diferentes sequências promotoras, ou de árvores sufixa, que representam as sequências promotoras em uma árvore sufixa. Ambas as abordagens tem o objetivo de encontrar número de frequência de uma subsequência, o qual deve ser comparado com o número de frequência esperada. Nesta etapa, são utilizados métodos estatísticos para avaliar a significância da sequência observada (ROMBAUTS et al., 2003).

Uma das primeiras abordagens de predição baseada em palavras é a Oligo-Analysis um algoritmo desenvolvido por Helden *et al*. (HELDEN; ANDRé; COLLADO-VIDES, 1998). O algoritmo desenvolvido foi utilizado na predição de elementos regulatórios em leveduras(*Saccharomyces cerevisiae*), as sequências promotoras utilizadas foram de genes co-regulados, e para o cálculo da frequência foi usado o método de enumeração exaustiva. O algoritmo mostrou-se eficiente na busca de elementos regulatórios pequenos e altamente conservados.

2.2.2 Algoritmos de predição probabilística

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

A introdução do EM para a busca de elementos regulatórios foi feita por Lawrence e Reilly (LAWRENCE; REILLY, 1990). O algoritmo assume que cada sequência tem pelo menos um elemento em comum. Diferente de outros algoritmos, neste algoritmo não é necessário o alinhamento das sequências.

Nos modelos de predição probabilística que utilizam técnicas de aprendizado de maquina, uma das técnicas que vem mostrando grande eficiência na busca de elementos regulatórios é a de *support vector machine* (SVM). Wang *et al.* (WANG et al., 2009) utilizaram sequências de DNA de regiões promotoras da *Arabidopsis*, que continham elementos regulatórios que se conectavam a TFs, de genes que eram ativados quando a planta é exposta ao estresse abiótico, e sequências aleatórias da região promotora. Foi aplicado então o algoritmo HexDiff (CHAN; KIBLER, 2005), este algoritmo encontra novos elementos regulatórios a partir de elementos já conhecidos, com elementos regulatórios encontrados ele então classificou-os com o SVM quais eram ativados em resposta ao estresse abiótico.

Hu *et al.* (HU; LI; KIHARA, 2005), propôs um algoritmo um algoritmo ensemble, que combina múltiplos algoritmos de identificação de elementos regulatórios. Os algoritmos usados foram AlingACE (ROTH et al., 1998) , Bioprospector (LIU; BRUTLAG; LIU, 2001), MDScan(LIU; BRUTLAG; LIU, 2002), MEME (BAILEY et al., 2006) e MotifSampler (THIJS et al., 2002). Este algoritmo segundo os autores,tem um grande desempenho, mas os resultados são mais precisos na identificação de elementos regulaórios pequenos.

2.2.3 Algoritmos baseados em rastros filogenéticos

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.

Blanchette *et al.* (BLANCHETTE; TOMPA, 2002), desenvolveram o FootPrinter. Este algoritmo que tem como entrada as sequências promotoras de várias espécies e a árvore filogenética das espécies relacionadas. As sequências de cada espécie são inseridas nas folhas da árvore, então são feitas varias comparações das sequências, a começar das folhas. As sequências de tamanho k, mais conservadas são promovidas para o nível acima da árvore, são feitas novas comparações até ser atingido a raiz da árvore, chegando a sequências ótimas, que passarão por mais comparações, desta vez da raiz até as folhas, finalizando o algoritmo com os elementos preditos.

2.2.4 Algoritmos que combinam técnicas probabilísticas e de rastros filogenéticos

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).

Sinha *et al.* (SINHA; BLANCHETTE; TOMPA, 2004), propuseram um algoritmo combinando as técnicas probabilísticas e de rastros filogenéticos. O algoritmo permite a entrada de sequências promotoras de genes ortólogos, relacionadas com a árvore filogenética definida pelo usuário. As sequências dos elementos regulatórios podem ser conservadas ou não conservadas, o algoritmo as trata de maneira diferente. O algoritmo permite uma flexibilidade, na escolha da árvore filogenética, podendo escolher espécies distantemente relacionadas, que além da identificação dos elementos conservados entre as espécies possibilita a identificação de elementos que não estão relacionados com as especies ortólogas.

Na tabela 1, estas listados alguns dos algoritmos dos modelos citados.

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 1: Implementações de modelos de predição

3 OBJETIVOS

Os objetivos gerais deste trabalho são:

- Identificar elementos regulatórios na soja ativados em condições de estresse abiótico.

Os objetivos específicos são:

- Desenvolver um algoritmo de classificação de elementos regulatórios ativados pelos fatores de transcrição da classe DREB.

4 PROCEDIMENTOS METODOLÓGICOS/MÉTODOS E TÉCNICAS

Para o desenvolvimento deste trabalho, serão realizadas as seguintes atividades:

- Levantamento bibliográfico.

  O levantamento bibliográfico, ocorrerá durante todo a realização do trabalho.

- Estudo de métodos para a predição de elementos regulatórios.

  Nesta fase serão estudas técnicas de predição de elementos regulatórios, identificando as técnicas com resultados mais precisos.

- Definição do conjunto de dados.

  Neste passo, será feita a coleta e separação de sequências promotoras do DNA do genoma da soja. Para isto, serão utilizados bancos de dados publicos que contêm sequências promotoras.

- Identificação e análise de elementos regulatórios.

  Nesta etapa, será implementado um sistema de identificação e classificação dos elementos regulatórios, que são ligados com fatores de transcrição da classe DREB.

- Validação dos elementos encontrados no genoma da soja.

  Neste ponto os elementos encontrados serão comparados com elementos já existentes, avaliando a precisão dos métodos implementados.

- Redação do TCC.

  O TCC será redigido durante todo a elaboração do trabalho.

5 CONTRIBUIÇÕES E/OU RESULTADOS ESPERADOS

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 (WANG; HABERER; MAYER, 2009) e um passo fundamental na construção da rede de regulação de um gene. Esse conhecimento é fundamental para interpretar e modelar as respostas de uma célula a diversos estímulos (WASSERMAN; SANDELIN, 2004).

Na soja, que é uma cultura amplamente usada na alimentação mundial, são poucos os estudos computacionais direcionados para a predição dos elementos regulatórios. Existem algumas ferramentas para a identificação dos elementos regulatórios em plantas e bancos de dados como o PLACE (HIGO et al., 1999), AGRIS (PALANISWAMY et al., 2006) e o PlantCARE (ROMBAUTS et al., 1999), mas a maior parte dos dados contidos neles, são referentes a planta Arabidopsis.

Com a descoberta de novos elementos regulatórios na soja e fatores de transcrição, é possível desenvolver através de engenharia genética, plantas tolerantes a estresses. Aumentando a produtividade e a qualidade dos grãos, assim como evitar o uso de agrotóxicos na cultivação da cultura.

6 CRONOGRAMA DE DESENVOLVIMENTO

O cronograma de desenvolvimento dos trabalhos será:

Tabela 2: Comparação das características dos métodos de clusterização estudados

Mês	2	3	4	5	6	7	8	9	10	11
Levantamento bibliográfico	X	X	X	X	X	X	X	X	X	
Estudo de métodos para a predição de elementos regulatórios	X	X	X	X						
Definição do conjunto de dados				X						
Identificação e análise de elementos regulatórios					X	X	X	X	X	
Validação dos elementos encontrados no genoma da soja									X	
Redação do TCC	X	X	X	X	X	X	X	X	X	X

Referências

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>.

ALBERTS, B. *Molecular biology of the cell*. 4. ed. Garland Science, 2002. Hardcover. ISBN 0815332181. Disponível em: <http://www.amazon.com/exec/obidos-/redirect?tag=citeulike07-20\&path=ASIN/0815332181>.

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>.

BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. *Biochemistry*. Sixth edition. W. H. Freeman & Co Ltd, 2007. Hardcover. ISBN 0716787245. Disponível em: <http://www.amazon.com-/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0716787245>.

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>.

HIGGS, P. G.; ATTWOOD, T. K. *Bioinformatics and molecular evolution*. Blackwell Pub., 2005. ISBN 9781405106832. Disponível em: <http://www.worldcat.org/isbn-/9781405106832>.

HIGO, K. et al. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic acids research*, Department of Genetic Resources, National Institute of Agrobiological Resources, 2-1-2 Kannondai, Tsukuba,Ibaraki 305, Japan. kenhigo@abr.affrc.go.jp, v. 27, n. 1, p. 297–300, jan. 1999. ISSN 0305-1048. Disponível em: <http://www.ncbi.nlm.nih-.gov/pmc/articles/PMC148163/>.

HU, J.; LI, B.; KIHARA, D. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, Oxford University Press, v. 33, n. 15, p. 4899–4913, set. 2005. ISSN 0305-1048. Disponível em: <http://dx.doi.org/10.1093/nar/gki791>.

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>.

LAWRENCE, C. E.; REILLY, A. A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Genetics*, Biometrics Laboratory, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201, v. 7, n. 1, p. 41–51, 1990. ISSN 0887-3585. Disponível em: <http://dx.doi.org/10.1002-/prot.340070105>.

LIU, X.; BRUTLAG, D. L.; LIU, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, Stanford Medical Informatics, 251 Campus Dr. X215, Stanford University, Stanford, CA 94305-5479, USA. xliu@smi.stanford.edu, p. 127–138, 2001. ISSN 1793-5091. Disponível em: <http://view.ncbi.nlm.nih.gov/pubmed/11262934>.

LIU, X. S.; BRUTLAG, D. L.; LIU, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology*, Stanford Medical Informatics, Stanford University, Stanford CA 94305, USA., v. 20, n. 8, p. 835–839, ago. 2002. ISSN 1087-0156. Disponível em: <http://dx.doi.org/10.1038/nbt717>.

PALANISWAMY, S. K. et al. AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant physiology*, Human Cancer Genetics Program, Comprehensive Cancer Center, Department of Molecular Virology, Immunology and Medical Genetics , The Ohio State University, Columbus, 43210, USA., v. 140, n. 3, p. 818–829, mar. 2006. ISSN 0032-0889. Disponível em: <http://dx.doi.org/10.1104/pp.105.072280>.

ROMBAUTS, S. et al. PlantCARE, a plant cis-acting regulatory element database. *Nucleic acids research*, Laboratorium voor Genetica, Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), Universiteit Gent,K.L. Ledeganckstraat 35, B-9000 Gent, Belgium., v. 27, n. 1, p. 295–296, jan. 1999. ISSN 0305-1048. Disponível em: <http://dx.doi.org/10-.1093/nar/27.1.295>.

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>.

WANG, X.; HABERER, G.; MAYER, K. F. X. Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. *BMC Genomics*, v. 10, n. 1, p. 284+, 2009. ISSN 1471-2164. Disponível em: <http://dx.doi.org/10.1186/1471-2164-10-284>.

WASSERMAN, W. W.; SANDELIN, A. Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics*, Nature Publishing Group, v. 5, n. 4, p. 276–287, abr. 2004. ISSN 1471-0056. Disponível em: <http://dx.doi.org/10.1038/nrg1315>.

ZAHA, A. *Biologia molecular básica*. Mercado Aberto, 2000. ISBN 8528002837. Disponível em: <http://www.worldcat.org/isbn/8528002837>.