TCC

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

Sumário

1 Introdução

Com o avanço na engenharia genética, hoje é possível criar plantas resistentes a algumas pragas e a fatores adversos a sobrevivência da planta. Mas para que isto ocorra é importante o entendimento da funcionalidade de cada gene no organismo e quais são as respostas do gene a um determinado estímulo. Ainda existem muitos genes com a funcionalidade não descoberta em diversos organismos, deixando um lacuna para novas pesquisas e metodologias que visão a descoberta da funcionalidade destes genes. Ao decorrer dos anos muitas fatores de transcrição foram descobertos assim como a funcionalidades de muitos genes, com isto muitas metodologias *in silico* surgiram aproveitando estas descobertas.

O problema de encontrar a funcionalidade de genes através de dados de genes e fatores de transcrição conhecidos, remete a problemas de classificação. Os algoritmos de aprendizado de máquina são conhecidos por obterem resultados bons com dados de difícil separação e alta dimensionalidade como são os dados genômicos. Esta revisão apresenta uma breve explicação dos fatores de transcrição com foque no *dehydration responsive element binding proteins*, também são apresentado métodos computacionais que auxiliam inferir a funcionalidade de um gene.

2 O papel do dehydration responsive element binding proteins (DREB) na transcrição

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 estresse abiótico afeta diversos organismos, mas em especial os organismos vegetais que são dependentes de fatores ambientais, são os mais afetados. O DREB faz parte do conjunto de fatores de transcrição relacionados a estímulos abióticos, e se destaca, porque ele regula genes que respondem aos estímulos abióticos de mudanças de temperatura, alta salinidade, e seca. Nas próximas seções são apresentadas as principais características relevantes para estudos computacionais, dos fatores de transcrição, elementos regulatórios e por ultimo o DREB.

2.1 Fatores de transcrição

Os fatores de transcrição (TFs, do inglês *Trascription Factors*), representam uma classe especial de proteínas (WERNER, 2009). Eles atuam no inicio da expressão de um gene, e operam na regulação da transcrição de um gene. Os TFs se conectam em pequenas regiões (5 a 20 nucleotídeos) no DNA chamadas de elementos regulatórios , quando conectados a um gene atuam na regulação do mesmo. A regulação pode ser tanto positiva (irá reforçar a transcrição), como negativa (irá inibir a transcrição). A regulação da transcrição de um gene é centralizada na expressão de um gene para um tecido específico (p. ex. um tecido de um órgão), e na regulação da ativação de um gene em resposta a um estimulo específico (LATCHMAN, 1997).

Pesquisas como a de (DAVIDSON; JACOBS; BRITTEN, 1983), mostraram que genes que são expressos em repostas a um determinado estímulo, como elevadas temperaturas, na região promotora destes genes é possível encontrar elementos regulatórios comuns entre estes genes. Mas estes elementos regulatórios não apareciam em genes que não respondiam ao

estímulo. Com esta especificidade dos elementos regulatórios, e consequentemente dos TFs como já comentado, é possível classificar TFs segundo os genes que eles regulam, de fato existem várias famílias de TFs em diversos organismos.

2.2 Elementos regulatórios

Os elementos regulatórios são pequenos seguimentos de sequência (5 a 20 nucleotídeos), localizados na região promotora[1] de um gene, geralmente são encontrados aproximadamente a uma distância de -50 pb do sítio de início da transcrição[2]. São nos elementos regulatórios que os fatores de transcrição irão se conectar.

Múltiplos elementos regulatórios formam os CRMs (do inglês, *cis-elements modules*), que integra a conexão de vários TFs resultando em um controle combinatório, e em um padrão específico da expressão de um gene (PRIEST; FILICHKIN; MOCKLER, 2009). Entendendo as funções dos elementos regulatórios e dos CRMs é essencial para compreender as respostas celulares ao ambiente (PRIEST; FILICHKIN; MOCKLER, 2009).

2.3 Dehydration responsive element binding proteins (DREB)

O *dehydration responsive element binding proteins* (DREB) é um importante fator de transcrição encontrado em plantas. O DREB ativa genes que estão relacionados com a resposta da célula a estímulos abióticos, com a ativação destes genes, a planta se adapta as condições adversas a sua sobrevivência, através de reações bioquímicas e físicas que ocorrem na planta. Os estresses abióticos que mais afetam as plantas são: seca, alta salinidade e mudanças de temperatura. O estresse abiótico atrapalha a sobrevivência e consequentemente a produção de grãos como a soja, arroz, milho e o trigo.

O DREB está contido dentro de uma família de fatores de transcrição única nas plantas, a *Ethylene Responsive Element* (ERF). A ERF desempenha um importante papel em resposta a estímulos abióticos e bióticos. O DREB pode ser dividido em duas subclasses DREB1 e DREB2, envolvidas em estresses de baixa temperaturas e desidratação, respectivamente (CHEN et al., 2007).

O elemento regulatório que se conecta no DREB é chamado de *dehydration res-*

[1] É uma região não transcrita do gene localizada antes da parte que é transcrita

[2] Ponto exato onde inicia-se a transcrição é um delimitador entre a região promotora e a parte onde vai ser transcrita

ponsive element DRE, e ele é formado pela sequência A/GCCGAC (NAKASHIMA; ITO; YAMAGUCHI-SHINOZAKI, 2009). Entretanto para que o DRE seja funcional ele tem que estar acompanhado por outros elementos regulatórios, mas a especificidade desses elementos regulatórios é baixa, fazendo que muitos desses varie de gene para gene (ZHANG et al., 2005). Portanto, em uma busca computacional de genes alvos do DREB (ou outro fator de transcrição), não basta procurar o DRE (ou outro elemento regulatório), na região promotora, uma vez que outros elementos regulatórios também participam da regulação, isto torna a busca por genes alvo uma difícil tarefa, uma vez que ha uma variação nos elementos regulatórios de um gene para outro.

Segundo (AGARWAL et al., 2006), o entendimento do DREB na regulação de um gene é de grande importância para o desenvolvimento de plantas tolerantes a estresses. Já que, 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.

3 Busca de genes alvos de TFs

Para descifrar a regulação genica de um organismo, e consequentemente ter um domínio na manipulação genética do organismo, muitas metodologias *in vitro* e *in silico* foram propostas. Das metodologias *in silico*, algumas focam na busca de genes alvos de TFs, e outras na classificação da funcionalidade de um gene desconhecido a partir de um gene conhecido. A seguir serão apresentados algumas metodologias para a classificação da funcionalidade de genes. Em especial a seção 3.1 apresenta a busca de genes alvos do DREB que é o foco desta pesquisa.

3.1 An in silico strategy identified the target gene candidates regulated by dehydration responsive element binding proteins (DREBs) in Arabidopsis genome (WANG et al., 2009)

Para inferir genes alvos do DREB na *Arabidopsis* (WANG et al., 2009), criaram uma estratégia computacional, que combina a análise de elementos regulatórios (TFBS, do inglês *Transcription factor binding site*) e aprendizado de máquina.

Eles utilizaram como conjunto de dados: sequências da região promotora (PR, do inglês *Promoter region*) de genes alvos do DREB identificados experimentalmente, estes são identificados como genes mestres (MGs, do inglês *Master genes*); *DRE frame sequences* (DFSs), fragmentos de DNA com 206 pb (pares de bases), retirados das PRs de MGs, contendo a subsequência consenso (A/GCCGAC) que é a região onde o DREB se conecta em um gene, identificada no artigo como DRE-motif; *Non-DRE frame sequences* (nDFS) fragmentos de de DNA com 206 pb, retirados das PRs de gene aleatórios, com um DRE-motif inserido artificialmente na região central. No total foram encontrados 48 DFSs de PRs de MGs, que foram considerados como dados positivos, e 1000 nDFSs como dados negativos.

Após a seleção do conjunto de dados, foi construído um classificador SVM (do inglês,

support vector machine) para categorizar DFSs e nDFSs. O vetor de características utilizado no SVM, foi um vetor contendo hexamers (pequenos fragmentos de DNA com 6 nucleotídeos), que foram selecionados através do algoritmo HexDiff (CHAN; KIBLER, 2005). Este algoritmo computa as pontuações da frequência de um hexamer no conjunto positivo $F_p(h)$ e negativo $F_n(h)$, em ambos filamentos de DNA. Depois de calculada a frequência para cada hexamer em ambos os conjuntos de dados foi calculado a pontuação da razão $R(h)$, como:

$$R(h) = \frac{F_p(h)}{F_n(h)} \tag{3.1}$$

Os hexamers com a pontuação da razão maior que um limite estabelecido, foram colocados em um vetor H_d, que é o vetor de características da SVM. Para cada hexamer em H_d que pertence ao conjunto de DFS foi atribuído o rótulo $(+1)$, para os de nDFS receberam (-1). A função Kernel utilizada no classificador SVM foi a RBF. Para o treinamento do classificador, visto que havia uma disparidade grande entre os dados positivos e negativos, foram usados apenas 100 amostras negativas e todas as 48 positivas. Já na classificação de DFSs e nDFSs foram utilizados as 1000 amostras negativas, os dados positivos foram reposicionados para que pudesse existir um sobre-amostragem, uma vez que eles estavam em um quantidade bem menor (48 DFSs), assim o conjunto de dados positivos passou para 500 amostras.

Para a previsão de genes alvos do DREB foi primeiramente, selecionados genes em todo o genoma da *Arabidopsis*, cuja a PR tem a sequência consenso do DREB (A/GCCGAC). Então foram selecionadas as regiões a esquerda da sequência consenso e a direita, ambas com 100 pb, que juntamente com o consenso forma uma subsequência de 206 pb. O conjunto de dados formado, com o procedimento descrito, foi classificado com o classificador SVM. Eram considerados genes alvos do DREB, todos os gene que contem pelo menos um DFSs em sua PR. No total, 474 genes foram preditos pela SVM e considerados fortes candidatos a serem alvos do DREB.

Segundo os autores, apenas encontrar regiões nas PRs de um gene contendo o consenso DREB, não é suficiente para inferir este gene como alvo do DREB, devido as perdas de características do DRE-motif. Vários estudos, apontam que TFBS distribuídos na PR influenciam na ligação de um TF e seu alvo. O que pode ser entendido que, na PR de genes alvos do DREB não somente o DRE-motif mas também outros TFBS, influenciam na conexão de um DREB na sequência. Estes outros TFBS agem como auxiliares para promover a conexão do DREB nos genes alvos. Com a utilização do HexDiff é selecionado regiões conservadas, que podem ser partes de TFBS que influenciam na ação do DREB, portanto um DFS será formado além da região consenso, também por sub-regiões conservadas.

3.2 Classifying transcription factor targets and discovering relevant biological features (HOLLOWAY; KON; DELISI, 2008)

Neste trabalho (HOLLOWAY; KON; DELISI, 2008), projetaram uma metodologia, utilizando aprendizado de máquina, para a predição de genes alvos de fatores de transcrição (TF, do inglês transcription factor) específicos. Com os resultados obtidos eles puderam construir e analisar a rede regulatória do *Saccharomyces cerevisiae*, uma levedura muito utilizada em estudos genéticos, por ter um genoma pequeno e mais simples comparado a outros organismos.

Para a classificação dos genes alvos de TFs, eles utilizaram máquinas de vetores de suporte (SVM, do inglês support vector machine). Como justificativa da utilização do SVM, os autores afirmam que: os conjuntos de dados genômicos têm uma alta dimensionalidade, detalhadamente, o conjunto pode chegar a milhares ou dezena de milhares de características numéricas para descrever um gene. Assim muitos algoritmos classificadores, podem ter um baixo desempenho com um numero alto de características, ao contrario do SVM que tem um bom desempenho com dados de alta dimensionalidade.

Como dados de entrada positivo, para o treinamento da SVM, foram pegos genes com elementos regulatórios conhecidos, que se ligam a determinados TFs, que também são conhecidos, sabe-se da existência dos elementos regulatórios, de seus respectivos genes e dos TFs, por meio de publicações na literatura. O conjunto negativo foi pego de subconjuntos de genes que têm um alto p-valor, consequentemente com uma probabilidade baixa de ter uma ligação aos TFs utilizados. Podemos perceber que, para cada TF em que deseja-se encontrar os genes que eles regulam é necessário a construção de um classificador (ou treinamento do classificador baseado no TF).

Para o treinamento do classificador, é recomendado que o conjunto negativo tenha pelo menos três vezes o tamanho do conjunto positivo. No conjunto negativo estão os genes que mostraram uma probabilidade baixa de serem regulados pelos TFs usados para prever os genes que eles regulam. Depois de selecionado o conjunto negativo são construídos 50 classificadores, para cada TF, utilizando diferentes subamostras do conjunto negativo, com o mesmo tamanho da amostra positiva. É utilizado 50 classificadores para cada TF, porque genes com alto p-valor são associados ao conjunto negativo e eles têm grandes probabilidades de não terem ligações com um dos TFs utilizados, mas há uma pequena probabilidade de um gene ser associado incorretamente ao conjunto negativo. Com os 50 classificadores este

inoportuno é suavizado.

Para a avaliação de cada classificador é usado a abordagem *leave-one-out cross-validation* (LOOCV), e também as medidas de desempenho: precisão e valor preditivo positivo (PPV, do inglês positive predictive value), que são usadas como média entre os 50 classificadores.

As características do classificador são selecionadas aplicando o algoritmo SVM *recursive feture elimination* (SVM-RFE), que otimiza o vetor **w** da SVM, para conter componentes altas, que são melhores para separar as classes positivas e negativas de dados. O processo de seleção de características, usando o SVM-RFE, é repetido até atingir o numero desejado de características, que é 1500, este numero é escolhido porque, quando é usado 1500 características a medida de precisão é aproximadamente 85%, que é uma precisão consideravelmente boa. Estas características são escolhidas para cada TF e são guardadas durante a avaliação dos 50 classificadores. No final haverá um grande conjunto de características para um TF, lembrando que os elementos desse conjunto são subconjuntos de características. Então são escolhidos os 40 maiores características, de cada subconjunto de características, e são guardas em uma lista, e é contada quantas vezes cada característica apareceu. Esta lista é rearranjada, posicionando os elementos que tem uma frequência de aparição maior no topo, assim este é o conjunto final de características. Essas características inclui um conjunto diverso de dados incluindo sequências promotoras, medidas de expressão de um gene, conservação filogenética de elementos de sequências, sobre-representação de sequências promotoras, temperatura de fusão de promotores, e outras.

O esquema usado para a montagem dos classificadores funciona da seguinte maneira: passo 1, é reunido o conjunto de dados positivos e negativos, totalizando *n*; passo 2, utilizando o LOOCV é pego *n-1* genes para o conjunto de treinamento e 1 para o conjunto de teste; passo 3 então é usado o SVM-RFE para classificar as características no conjunto de treinamento; passo 4 construir um classificador SVM com as 1500 características. Salvar as características; passo 5, classificar o gene deixado de fora do conjunto de treinamento; passo 6 repetir os passos 2-5 até completar o LOOCV. Salvar todas as características; passo 7 calcular as estatísticas de desempenho (precisão, PPV, etc.); passo 8 repetir passos 1-5 50 vezes; último passo, calcular as estatísticas de desempenho final (média da precisão, média do PPV).

Os autores aplicaram este método em 163 TFs do *S. cerevisiae*, com os resultados obtidos eles construíram uma rede de regulação que foi disponibilizada em um servidor web (http://cagt10.bu.edu/TFSVM/main.htm), onde é possível consultar quais são os genes alvos de um TF, ou quais TFs se ligam em um gene.

3.3 Combining classifiers to predict gene function in Arabidopsis thaliana using large-scale gene expression measurements (LAN et al., 2007)

O foco deste trabalho é a classificação dos genes segundo suas funcionalidades, mais especificamente de genes que estão envolvidos na resposta das plantas a estresses. O trabalho foi conduzido por (LAN et al., 2007), e a planta usada como teste foi a *Arabidopis thaliana*.

Para a classificação dos genes foram desenvolvidos cinco métodos de aprendizado supervisionado, que foram: *Logistic Regression* (LR), *Linear Discriminant Analysis* (LDA), *Quadratic Discriminant Analysis* (QDA), *Naive Bayes* (NB) e *K-Nearest Neighbors* (KNN). Foram escolhidos estes métodos básicos, porque eles requerem pouca computação e os resultados são bons o suficiente para se fazer análises biológicas. Todos os classificadores montados com os métodos, retornam um valor discriminativo para cada gene avaliado. Cada gene é representado por vetor com 290 dimensões cujas as componentes são valores de expressão do gene em 290 condições experimentais. O valor retornado por um classificador tem que ser maior que o limite estabelecido.

O conjunto de dados usado foi extraído de experimentos relacionados com estresses. No total foram 22.746 genes sobre 290 condições experimentais diferentes. Desses 22.746 genes, 11.553 são genes anotados e com suas funções conhecidas, e desses 1.031 respondem à estresses. Os genes anotados formaram o conjunto de dados de treinamento, onde um gene foi considerado positivo se ele foi anotado como um gene de resposta a estresse, e negativo para os demais. Foram 11.193, o total de genes não anotados, estes foram usados para fazer as previsões. Podem haver alguns falsos negativos, visto que genes que não foram descobertas as funcionalidades, são introduzidos no conjunto negativo, entretanto eles podem ser genes de resposta a estresses.

Antes dos dados serem treinados, eles passaram por um pré-processamento, para reduzir a dimensão do conjunto de dados, para que durante o aprendizado supervisionado os dados sejam usados eficientemente. Isto foi feito com o algoritmo *Principal components analysis* (PCA), que mapeia vetores de alta dimensão para um dimensão menor. O PCA aplicado ao conjunto de dados, que originalmente tinha 290 dimensões, reduziu a dimensão para as dimensões de 5, 10, 15, 20, 40 e 100.

Os classificadores são treinados com todas as dimensões geradas pelo PCA e pela dimensão original com exceção do KNN, no caso do KNN é usado diferentes valores de K na dimensão original. Então é escolhida a dimensão em que cada algoritmo obteve melhores

resultados (no caso do KNN o melhor K). Os classificadores com a sua melhor dimensão, são combinados em um só classificador, onde o valor de discriminação do classificador combinado é uma combinação linear dos discriminastes dos classificadores individuais. Como esperado o classificador combinado obteve melhores resultados que os classificadores individuais.

O resultado final da classificação é um conjunto de genes, que podem ser posicionados quanto ao valor discriminante do gene responder a estresses, ficando os com maiores valores no topo.

3.4 Using hexamers to predict cis-regulatory motifs in Drosophila (CHAN; KIBLER, 2005)

Neste trabalho (CHAN; KIBLER, 2005) desenvolveu o algoritmo HexDiff. Este algoritmo busca agrupações de TFBS, que atuam juntos na regulação de um gene. Um agrupamento de TFBS é comumente conhecido por CRM (do inglês *cis-regulatory modules*).

O HexDiff é um tipo de algoritmo de aprendizado de máquina, e foi projetado para discriminar dois tipos de sequências de DNA: CRM, e non-CRM(não agrupamento de TFBS). Para fazer a classificação é necessário um conjunto de dados de treinamentos, que é obtido através de conhecidos CRMs, que são colocados no conjunto positivo de treinamento, os não conhecidos, os non-CRMs, são inseridos no conjunto negativo. Os dados para teste foram pegos de citar, de conhecidos CRMs da *Drosophila*. Foram encontrados 16 genes que, contém um total de 52 CRMs. Após a seleção dos dados é calculada a frequência de cada hexamer (subsequência de nucleotídeos de 6 pb), no conjunto negativo $f_p(h)$ e positivo $f_n(h)$, para calcular a razão $R(h)$:

$$R(h) = \frac{F_p(h)}{F_n(h)} \tag{3.2}$$

Os hexamers que obtiverem um alto valor de $R(h)$, são colocados no conjunto H_d. Com isto H_d vai ter hexamers que são mais comuns em CRMs do que em non-CRMs.

Depois de gerado o conjunto H_d, ele é usado para classificar cada posição em uma sequência não conhecida como uma sequência CRM e non-CRM. Para fazer a classificação é construído uma janela na sequência, entre 1000-2000 pb, que a cada rodada é movida 1 pb na sequência, e é calculado a pontuação S_i para cada posição i da janela na sequência, pelo produto da razão $R(h)$ e o numero de aparições de um hexamer $n(h_d)$ em H_d na janela, qualquer posição que exceder o limite é considerada um CRM.

Para a avaliação foi usado a abordagem *leave-one-out cross-validation* (LOOCV),

onde dos 16 genes encontrados 15 são treinados e 1 é usado como teste, este processo é feito até que todos os 16 genes sejam usados no conjunto de teste. A precisão das previsões do modelo foi medida com a correlação de Matthew.

Quando aplicado a no-CRMs da *Drosophila*, além dos CRMs já conhecidos serem previstos, outros 10 CRMs foram encontrados e indicados como fortes candidato a CRMs.

3.5 Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana (ZHANG et al., 2005)

Neste trabalho os autores, projetaram uma aplicação para para encontrar genes alvos de fatores de transcrição, na *Arabidopsis thaliana* ,que são induzidos por ácido abscísico(ABA, do inglês *abscisic acid*)e por estresses abióticos.

Os dados utilizados foram coletados de plantação de *A.thaliana*, que cresceu a uma temperatura de 24°C durante 10 dias, algumas mudas foram expostas ao ABA e/ou estresses abióticos, e outras não. Das sequências extraídas do experimento, foram coletados no total 366 regiões promotoras de genes regulados após a exposição ao ABA e/ou estresses abióticos para análises.

Para encontrar os genes alvos, primeiramente foi calculado a pontuação de cada subsequência de tamanho w em uma determinada sequência, baseado em padrões com relevâncias biológicas (*motifs*) e um modelo de *background*. Um *motif* W de tamanho w é representado por uma matriz de peso (PMW, do inglês *Position Weight Matrix*), com $PWM\Theta_W = (q_{l,b})$, onde $(q_{l,b})$ é a probabilidade de encontrar a base b na posição i do *motif*, já o modelo de *background* é criado utilizando o modelo de Markov. O modelo de *background* calcula a probabilidade de uma base começar na $j-$ésima posição com $P(j|B_m) = \prod_{l=1}^{w} P(b_{j_l-1}|b_{j+l-2}...b_{j+l-l})$, onde B_m é a $m-$ésima ordem do modelo de Markov, e b_j é a $j-$ésima base da sequência. Também é calculada a probabilidade de uma subsequência ser um *motif* Θ_W, utilizando também o modelo de Markov, com $P(j|\Theta_W) = \prod_{l=1}^{w} q_{l,b_{j+l-1}}$, onde $q_{l,b_{j+l-1}}$ é a probabilidade de encontrar a base b_{j+l-1} na posição l-ésima de um *motif*. Então foi calculada a *log-ratio* $A_{j,\Theta_W,B_m} = ln\frac{P(j|\Theta)_W}{P(j|B_m)}$. A pontuação de uma sequência S é feita utilizando dois *motifs* Θ_M e Θ_N , neste caso um deles é o ABRE que é o elemento regulatório alvo, quando uma planta é exposta ao ABE, são consideradas todas as posições i e j dentro de S e a combinação da pontuação das posições é computada como $A_{S,\Theta_M,\Theta_N,B_m} = max_{i,j}(A_{i,\Theta_M,B_m} + A_{j,\Theta_N,B_m})$. A

combinação com maior pontuação é atribuída para a sequência, genes com sua sequência com maior pontuação têm uma probabilidade maior de serem genes alvos.

Como resultados foram encontrados entre 1825 genes induzidos a estresses, 1530 onde pelo menos uma funcionalidade poderia ser atribuída. E foram selecionados 150 com maior pontuação onde 126 estão classificados em alguma categoria funcionalidade. O que levou aos autores a concluir que podem existir muitas atividades de regulações genicas após a exposição ao ABA.

3.6 CisOrtho: A program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting (BIGELOW et al., 2004)

CisOrtho é software desenvolvido por (BIGELOW et al., 2004), que identifica alvos de fatores de transcrição com um específico elemento regulatório definido, utilizando rastros filogenéticos. O programa foi usado nos genomas de dois invertebrados, o *Caenorhabditis elegans* e o *Caenorhabditis briggsae*.

O primeiro passo é, identificar, classificar, e associar as regiões não codificantes dos genes, isto é feito com as informações contidas em arquivos GFF (do inglês, *General Feture Format*). Foram retiradas as regiões codificantes porque é improvável que exita elementos regulatórios nestas regiões e por elas serem improprias para buscas filogenéticas apresentando uma alta conservação. O segundo passo consiste em construir uma PWM (do inglês, *Position Weight Matrix*) para um conjunto de elementos regulatórios definidos experimentalmente, neste trabalho foram usados os elementos regulatórios dos fatores de transcrição TTX-3 e CEH-10. Para a construção da PWM foi usado o software HMMER (EDDY, 1998), este software utiliza o modelo oculto de Markov para gerar a PWM. A PWM resultante é usada como entrada no CisOrtho, que faz uma busca nas sequências não codificantes, atribuindo uma pontuação para cada bloco de subsequência de tamanho n (chamado de janela), onde o objetivo é encontrar subsequências com a maior pontuação N, e com no máximo D subsequências, onde N e D são definidos pelo usuário. No quarto passo é feita a análise filogenética, são utilizados conjuntos de mapeamentos ortólogos[1] entre *C. elegans* e *C. briggasae* que prove pares de subsequências ortólogas. Os pares de subsequências que apresentam uma alta pontuação são guardados, como um par de subsequência. O ultimo passo classifica os pares de subsequências

[1]Sequências conservadas de diferentes organismo, que tem o mesmo ancestral

de acordo com as suas pontuações e *mismatches*[2].

Com o CisOrtho foi possível identificar novos genes alvos dos fatores de transcrição TTX-3 e CEH-10, foram encontrados 14 genes com subsequências com alta pontuação que satisfaziam os critérios para serem alvos de TTX-3/CEH-10, também foram encontrados genes com baixa pontuação, mas que atendiam os critérios para serem alvos de TTX-3/CEH-10, um total de 11 genes. Para subsequência com uma grande conservação mas, com uma baixa pontuação, foram feitas análises e foram encontrados também genes alvos.

[2]Bases diferentes, em uma mesma posição, neste caso é comparado as bases do *C. elegans* e *C. briggasae*

3.7 Tabela comparativa

Trabalho	Organismo	Entradas	Técnicas usadas	Resultados obtidos
(WANG et al., 2009)	*Arabidopsis*	sequências promotoras de genes que contém o DRE-motif e sequências que não contém o DRE-motif	O algoritmo HexDiff e SVM	474 genes alvos
(HOLLOWAY; KON; DELISI, 2008)	*Saccharomyces cerevisiae*	sequências promotoras; medidas de expressão de um gene; conservação filogenética de elementos de sequências; sobre-representação de sequências promotoras; temperatura de fusão de promotores; K-mer conservados; k-mer com *mismatches*; *k-mer median position*	SVM	rede de regulação do *Saccharomyces cerevisiae*

Trabalho	Organismo	Entradas	Técnicas usadas	Resultados obtidos
(LAN et al., 2007)	*Arabidopis thaliana*	22.746 genes sobre 290 condições experimentais diferentes	*Logistic Regression* (LR), *Linear Discriminant Analysis* (LDA), *Quadratic Discriminant Analysis* (QDA), *Naive Bayes* (NB) e *K-Nearest Neighbors* (KNN)	Genes com grandes possibilidades de serem expressos mediante a condições de estresses abióticos
(CHAN; KIBLER, 2005)	*Drosophila*	CRM e não-CRM	HexDiff	10 novos CRMs
(ZHANG et al., 2005)	*Arabidopsis thaliana*	dados de plantas expostas ao ABA e/ou estresses abióticos	PWM e modelo de Markov	1530 genes que pode ser adicionada alguma funcionalidade
(BIGELOW et al., 2004)	*Caenorhabditis elegans* e o *Caenorhabditis briggsae*	arquivos GFF e conjuntos de mapeamentos ortólogos	Modelo oculto de Markov, PWM e rastros filogenéticos	mais de 25 genes alvos de TTX-3 e CEH-10

4 Conclusão

Podemos observar que a maioria dos trabalhos realizados na busca de genes alvos de fatores de transcrição, estão em comum acordo quanto a degeneração dos elementos regulatórios, e a dependência de um elemento regulatório de outros elementos regulatórios para a regulação de um gene formando os CRMs. Essas observações foram comprovadas em experimentos *in vitro*, como visto em (DAVIDSON; JACOBS; BRITTEN, 1983),(PRIEST; FILICHKIN; MOCKLER, 2009) e (ZHANG et al., 2005).

Observamos o importante papel que algoritmos de aprendizado de máquina tem na busca de genes alvos, estes algoritmos são usados em pelo menos uma das etapas das metodologias apresentadas. Em (HOLLOWAY; KON; DELISI, 2008) e (LAN et al., 2007), é discutido que os dados genômicos apresentam uma alta dimensionalidade, o que nos leva a concluir que a predominância do SVM nas metodologias é devido a facilidade que SVM tem em lidar com alta dimensionalidades e os bons resultados obtidos mediante estas condições.

Quanto uma análise de desempenho e precisão entre os algoritmos fica inconclusiva, uma vez que os trabalhos foram feitos em diferentes tipos de organismos, mesmo os trabalhos em os organismo eram os mesmos, eles eram aplicados a diferentes tipos de fatores de transcrição.

Finalizando, é de grande importância descobrir os genes alvos de fatores de transcrição, uma vez que, ainda existem vários genes que ainda não foi associada nenhuma função. Como a função de um gene está diretamente ligada ao tipo de fatores de transcrição que conectam nele, é possível classificar estes genes a partir de fatores de transcrição conhecidos.

Referências Bibliográficas

AGARWAL, P. et al. Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. *Plant Cell Reports*, Springer Berlin / Heidelberg, v. 25, n. 12, p. 1263–1274, dez. 2006. ISSN 0721-7714. Disponível em: <http://dx.doi.org/10.1007/s00299-006-0204-8>.

BIGELOW, H. et al. CisOrtho: A program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. *BMC Bioinformatics*, v. 5, n. 1, p. 27+, 2004. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186/1471-2105-5-27>.

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>.

CHEN, M. et al. GmDREB2, a soybean DRE-binding transcription factor, conferred drought and high-salt tolerance in transgenic plants. *Biochemical and biophysical research communications*, v. 353, n. 2, p. 299–305, fev. 2007. ISSN 0006-291X. Disponível em: <http://dx.doi.org/10.1016/j.bbrc.2006.12.027>.

DAVIDSON, E. H.; JACOBS, H. T.; BRITTEN, R. J. Eukaryotic gene expression: Very short repeats and coordinate induction of genes. *Nature*, Nature Publishing Group, v. 301, n. 5900, p. 468–470, fev. 1983. Disponível em: <http://dx.doi.org/10.1038/301468a0>.

EDDY, S. R. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, Department of Genetics, Washington University School of Medicine, 4566 Scott Avenue, St Louis, MO 63110, USA. eddy@genetics.wustl.edu, v. 14, n. 9, p. 755–763, jan. 1998. ISSN 1367-4803. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/14.9.755>.

HOLLOWAY, D.; KON, M.; DELISI, C. Classifying transcription factor targets and discovering relevant biological features. *Biology Direct*, v. 3, n. 1, p. 22+, maio 2008. ISSN 1745-6150. Disponível em: <http://dx.doi.org/10.1186/1745-6150-3-22>.

LAN, H. et al. Combining classifiers to predict gene function in Arabidopsis thaliana using large-scale gene expression measurements. *BMC Bioinformatics*, v. 8, n. 1, p. 358+, set. 2007. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186/1471-2105-8-358>.

LATCHMAN, D. S. Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology*, v. 29, n. 12, p. 1305–1312, dez. 1997. ISSN 13572725. Disponível em: <http://dx.doi.org/10.1016/S1357-2725(97)00085-X>.

NAKASHIMA, K.; ITO, Y.; YAMAGUCHI-SHINOZAKI, K. Transcriptional Regulatory Networks in Response to Abiotic Stresses in Arabidopsis and Grasses. *Plant Physiology*, v. 149, n. 1, p. 88–95, jan. 2009. ISSN 0032-0889. Disponível em: <http://dx.doi.org/10-.1104/pp.108.129791>.

PRIEST, H. D.; FILICHKIN, S. A.; MOCKLER, T. C. cis-Regulatory elements in plant cell signaling. *Current Opinion in Plant Biology*, v. 12, n. 5, p. 643–649, out. 2009. ISSN 13695266. Disponível em: <http://dx.doi.org/10.1016/j.pbi.2009.07.016>.

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>.

WERNER, T. The Role of Transcription Factor Binding Sites in Promoters and Their In Silico Detection. In: KRAWETZ, S. (Ed.). *Bioinformatics for Systems Biology*. Totowa, NJ: Humana Press, 2009. cap. 17, p. 339–352. ISBN 978-1-934115-02-2. Disponível em: <http://dx.doi.org/10.1007/978-1-59745-440-7_17>.

ZHANG, W. et al. Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana. *Bioinformatics*, Oxford University Press, v. 21, n. 14, p. 3074–3081, jul. 2005. ISSN 1367-4803. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bti490>.