Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Identificação de elementos cis-reguladores (CREs) nos promotores de genes co-expressos em cana-de-açúcar

João Vitor Leite Novoletti

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do título de Bacharel e Licenciado em Ciências Biológicas

Piracicaba
2022

João Vitor Leite Novoletti

Identificação de elementos cis-reguladores (CRE) nos promotores de genes co-expressos em cana-de-açúcar

Orientador:
Prof. Dr. **DIEGO MAURICIO RIAÑO-PACHÓN**

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do título de Bacharel e Licenciado em Ciências Biológicas

Piracicaba
2022

AGRADECIMENTOS

À Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ).

Ao meu orientador Prof. Dr. Diego Mauricio Riaño-Pachón, por toda sua paciência e acompanhamento durante o processo de construção desse trabalho. Agradeço por oferecer essa oportunidade de aprofundamento na área da Bioinformática que me deu um norte no caminho para seguir.

A todos os professores que contribuíram com a minha jornada na Faculdade durante esses anos. Agradeço por todos os ensinamentos e vivências desenvolvidas.

A todos os funcionários, em especial aos porteiros e ao pessoal da limpeza, por sempre deixarem a Faculdade organizada e segura. Em especial à funcionária da Divisão de Biblioteca, que colaborou com orientações sobre a formatação do trabalho.

Aos meus pais Antônio Carlos Novoletti e Janusi Leite Novoletti, por me apoiarem nessa jornada e propiciarem a oportunidade de estudo, sempre me apoiando e lidando com as minhas frustrações. Muito obrigado!

À minha irmã Jaqueline Vitória Leite Novoletti por estar sempre presente para conversar e por sempre me apoiar com seus conselhos. Obrigado.

A todos os meus parentes e amigos que fizeram parte da minha jornada, os quais influenciaram na minha escolha pelo curso de Ciências Biológicas. Muito Obrigado!

“A persistência é o caminho do êxito”

Charles Chaplin, 1997.

SUMÁRIO

RESUMO

Identificação de elementos cis-reguladores (CRE) nos promotores de genes co-expressos em cana-de-açúcar

A cana-de-açúcar é uma planta de grande importância econômica para o país, sendo uma das principais culturas no valor da produção agrícola, justificando estudos que visam aumento em sua produção. O entendimento da regulação de seus genes pode gerar base para essa finalidade. Esse trabalho se propôs em avaliar elementos cis-reguladores (CREs) em promotores de genes co-expressos na folha de cana-de-açúcar, através da busca “de novo” de motivos e comparação com um banco de dados de elementos cis-reguladores (JASPAR), a fim de entender os fatores que atuam na regulação da transcrição desses genes. Houve a identificação de 2579 motivos que foram comparados entre si, agrupados por sua similaridade e comparados contra o JASPAR. Dessa comparação, 98 motivos obtiveram semelhança com os sítios de ligação de fatores de transcrição (FT) conhecidos. Em seguida, foi realizado análise dos FTs resultantes conforme sua família e outras subclassificações para inferência de informação funcional dos motivos. Foram identificados sítios de ligação com FTs pertencentes à várias famílias, dentre elas, AP2/ERF, MADS-Box, DOF, Myb-related, BBR/BPC, TCP, GRF, que apresentam funções no desenvolvimento, morfologia e resposta ao estresse. Concluiu-se que a estratégia de avaliação de motivos em genes co-expressos foi eficaz na avaliação de CREs em cana-de-açúcar, apesar de que não houve comprovação de especificidade de expressão a “clusters” determinados por meio da ferramenta utilizada. Além disso, foi apontada a necessidade da realização de estudos acerca de elementos cis-reguladores presentes em monocotiledôneas, objetivando ampliar a gama de dados disponíveis para comparação.

Palavras-chave: Regulação, Promotores, Elementos cis-reguladores, Genes co-expressos, Cana-de-açúcar, Motivos

ABSTRACT

Identification of cis-regulatory elements (CRE) in promoters of co-expressed genes in sugarcane

Sugarcane is a plant of great economic importance for the country, being one of the main crops in the value of agricultural production, justifying studies hat aim to increase its production. Understanding the regulation of their genes can generate a basis for this purpose. This work aimed to evaluate cis-regulatory elements (CREs) in promoters of co-expressed genes in the sugarcane leaf, through a “de novo*”* search for motifs and comparison with a database of cis-regulatory elements (JASPAR), in order to understand the factors that act in the regulation of the transcription of these genes. There was the identification of 2579 motifs that were compared among themselves, grouped by their similarity and compared against JASPAR. From this comparison, 98 motifs were similar to the transcription factor (TF) binding sites. Then, analysis of the resulting TFs was carried out according to their family and other subclassifications for inference of functional information of the motifs. Binding sites were identified with TFs belonging to several families, among them, AP2/ERF, MADS-Box, DOF, Myb-related, BBR/BPC, TCP, GRF, which have functions in development, morphology and stress response. It was concluded that the evaluation strategy of motifs in co-expressed genes was effective in the evaluation of CREs in sugarcane, although there was no proof of specificity of expression to clusters determined through the used tool. In addition, the need for studies on cis-regulatory elements present in monocotyledons was pointed out, aiming to expand the range of data available for comparison.

Keywords: Regulation, Promoters, Cis-regulatory elements, Co-expressed genes, Sugarcane, Motif

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO/JUSTIFICATIVA

1.1 Cenário econômico da cana-de-açúcar

A cana-de-açúcar é uma das principais culturas em valor de produção agrícola no Brasil, superada apenas pelo milho e pela soja (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE, 2022). No que se refere a sua produção, estima-se que na safra de 2022/23 seja produzido em torno de 572,9 milhões de toneladas, gerando 25,83 bilhões de litros de etanol e 33,89 milhões de toneladas de açúcar (COMPANHIA NACIONAL DE ABASTECIMENTO - CONAB, 2022), demonstrando sua importância tanto nos setores agrícolas quanto industriais dentro do cenário nacional. Destaca-se que a biomassa mais utilizada para a produção de eletricidade é derivada da cana-de-açúcar, representando 19,1% da oferta interna de energia e mais de um terço de toda a energia utilizada no Brasil e participando de 13,8% de renováveis na matriz em nível mundial (EMPRESA DE PESQUISA ENERGÉTICA - EPE, 2022).

No cenário atual de mudanças climáticas, existem muitos esforços para ampliar a oferta de fontes renováveis de energia, como a biomassa da cana-de-açúcar. Visto isso, há cada vez mais estudos visando o aumento e melhoria de sua produtividade (EMBRAPA, 2021). É cada vez mais claro que o futuro do melhoramento de culturas de importância econômica passa pela modificação dos componentes das redes de regulação da expressão gênica, ou seja, os fatores de transcrição e/ou os promotores alvos destes (GAO et al., 2019).

1.2 Informação genômica da cana-de-açúcar

As variedades comerciais modernas de cana-de-açúcar resultaram de um processo que envolveu a hibridização de várias espécies do gênero *Saccharum*, a maioria delas poliploides (SFORÇA, 2019). A cana-de-açúcar possui sequências genômicas para três dessas variedades até o momento que foram geradas por sequenciamento de leituras longas (GARSMEUR et al., 2018; RIAÑO-PACHÓN; MATTIELLO, 2017; SOUZA et al., 2019; TRUJILLO-MONTENEGRO et al., 2021). A variedade SP80-3280, que se trata do cultivar com maior quantidade de dados moleculares disponíveis em bancos de dados públicos, possui ao menos dois rascunhos ainda muito fragmentados, observando-se a alta quantidade de “contigs” e o curto tamanho do N50 (Tabela 1) (RIAÑO-PACHÓN; MATTIELLO, 2017; SOUZA et al., 2019). “Contig” é o nome dado ao fragmento montado por sobreposição de “reads”

(leituras de sequenciamento), já N50 representa o comprimento da sequência do “contig” mais curto a ser somado para cobrir ao menos 50% do comprimento total da montagem, essa soma é realizada do maior “contig” para o menor. Rossi (2022), em sua dissertação, organizou os genes dessa variedade em uma representação conjunta e não redundante, além de realizar um agrupamento por genes co-expressos baseando-se em um estudo de desenvolvimento da folha de cana-de-açúcar. No estudo de base em questão, realizado por Mattiello et al. (2015), houve coleta e análise de dados do transcriptoma da folha +1 (a primeira folha com a barbela da bainha totalmente exposta) de cana-de-açúcar, em quatro segmentos da folha, chamados de Base0, Base, Médio e Ponta, e que representam diferentes estágios do desenvolvimento da folha. Dentre os resultados desse grupo de trabalho, foram identificados componentes relacionados ao estresse abiótico e biótico e fatores de transcrição (FTs) associados ao desenvolvimento e polaridade da folha sendo expressos diferencialmente em segmentos distintos.

Tabela 1. Métricas das sequencias genômicas de híbridos de cana de açúcar

Fonte	Nº de “contigs”	N50 (kbp)
RIAÑO-PACHÓN; MATTIELLO (2017)	199,028	8,4
SOUZA et al. (2019)	450,608	13,2

Fonte: Riaño-Pachón e Mattiello (2017) e Souza et al. (2019)

1.3 Princípios de regulação da expressão gênica

O fluxo de informação na célula, nomeado originalmente como "Dogma Central" (CRICK, 1958) traz o genoma como base para a expressão gênica que engloba a síntese de RNA mensageiro, RNA ribossomal, RNAs transportadores e proteínas. Na regulação dos padrões da expressão gênica, pode-se identificar o papel central das regiões promotoras dos genes (JUVEN-GERSHON; KADONAGA, 2010). Conforme Thrall et al. (2013), a região promotora orienta a transcrição do gene, nesta região estão presentes elementos reguladores que recrutam FTs específicos; esses, por sua vez, controlam o nível de transcrição quando ligados ao DNA. Esses agentes recrutadores são denominados elementos cis-reguladores (CREs) e contribuem para

a diversidade nos padrões de expressão tanto temporalmente quanto espacialmente (WITTKOPP; KALAY, 2012). Durante a transcrição da grande maioria dos RNAs mensageiros, a atuação dos FTs, após sua ligação com seu respectivo CRE, envolve uma interação com a RNA polimerase II, seja ela direta ou através de um mediador (KORNBERG, 2007) (Figura 1).

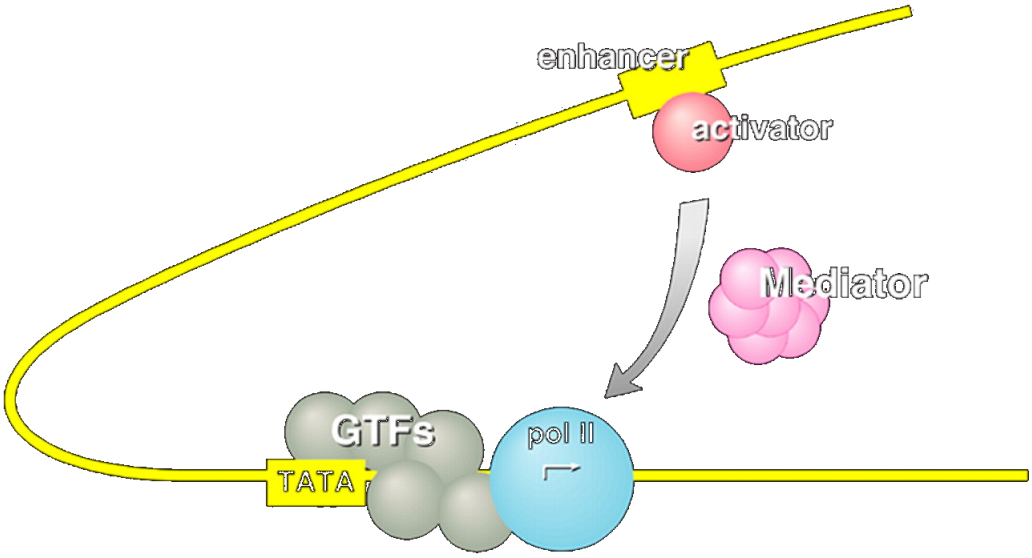

Figura 1. Representação esquemática dos processos regulatórios mediados pelos elementos presentes na região promotora do gene (KORNBERG, 2007).

Conforme Gonzalez (2016a), os FTs são proteínas detentoras de um domínio de ligação com o DNA que reconhece os CREs e agem como ativadores ou repressores, aumentando ou diminuindo a expressão do gene, geralmente, através da interação com outros componentes do complexo de início da transcrição. Isso ocorre devido à presença de outros domínios de interação presentes nos FTs localizados em regiões distintas, e de ação independente, em relação ao domínio de ligação com o DNA (GONZALEZ, 2016a).

No que concerne à presença desses reguladores em plantas, apesar de muitos FTs serem compartilhados com outros grupos de eucariotos, há a presença de alguns mais específicos a esse grupo (GONZALEZ, 2016a). Isso torna o estudo dos fatores de transcrição em plantas necessário para entender sua função e especificidade.

1.4 Ferramentas para análise de promotores e identificação de CREs

Há vários métodos de análise de promotores que se diferem na estratégia e dados utilizados. Em relação ao descobrimento de novos CREs, esse pode ser

realizado através de técnicas de estudo “de novo” que buscam por “motivos” nas regiões promotoras de genes co-expressos (AERTS, 2012). Motivos são sequências curtas e conservadas de nucleotídeos e apresentam sítios putativos de ligação proteica, sendo sua descoberta importante para o entendimento da regulação da transcrição (ZAMBELLI; PESOLE; PAVESI, 2012).

O STREME (BAILEY, 2021) integrante do “MEME suite” (BAILEY et al., 2015) é um dos programas que podem ser utilizados para realizar a descoberta de motivos fazendo uso de uma estrutura de dados chamada árvore de sufixos generalizada e pontuando matrizes de peso de posição (PWMs). Para avaliar os motivos, o STREME utiliza um teste estatístico unilateral do enriquecimento de combinações para o motivo em um conjunto de sequências alvo em comparação com um conjunto controle (BAILEY, 2021). Os possíveis sítios de ligação são modelados PWMs, definindo uma pontuação para cada um dos quatro nucleotídeos possíveis para cada posição do motivo (WEINER, 1973). Esse programa também apresenta um processo de refinamento do motivo em que uma sequência é revisada até a seleção daquela que melhor discrimina as sequências alvo das sequências controle (BAILEY, 2021). Estudos recentes em sequências em tandem no genoma humano (NISSANI; ULITSKY, 2022) e em redes de co-expressão em *Camellia sinensis* (ZHENG et al., 2022) utilizaram essa ferramenta, por exemplo.

Dentre as entradas dos programas de descobrimento de motivos, dados de co-expressão de genes são frequentemente utilizados (AERTS, 2012; MISHRA, 2018), visto que uma das razões para que um grupo de genes seja co-expressado é a coordenação em seu processo regulatório, sugerindo que as regiões promotoras compartilhem CREs (LIU; LI; CHENG, 2019). Em cana-de-açúcar, por exemplo, um estudo com a variedade SP80-3280 identificou motivos em promotores de genes co-expressos durante estresse hídrico e analisou FTs que, possivelmente, se ligam a eles (DINIZ et al., 2020). Dessa forma, estudos afins são chave para discriminar os mecanismos de regulação das plantas em diferentes fases de seu desenvolvimento.

Assim, neste trabalho buscou-se responder à pergunta: É possível identificar CREs putativos nos grupos de genes co-expressos em cana-de-açúcar durante o desenvolvimento da folha? E, em caso afirmativo, é possível identificar FTs que potencialmente reconhecem esses CREs?

.

2 OBJETIVOS

2.1 Objetivos gerais

Identificar elementos cis-reguladores em promotores de genes co-expressos durante o desenvolvimento da folha em cana-de-açúcar através de análise *in sílico*.

2.2 Objetivos específicos

Identificar motivos em promotores de genes co-expressos em cana-de-açúcar.

Verificar a ocorrência dos motivos em bancos de dados.

Analisar os fatores de transcrição que se ligam aos motivos encontrados.

Verificar a recorrência de elementos cis-reguladores nos promotores da cana-de-açúcar.

3 METODOLOGIA

3.1 Dados de genes co-expressos

Foram usados dados provenientes da dissertação de mestrado de Rossi (2022), que identificou 99 grupos (“clusters”) de genes co-expressos durante o desenvolvimento da folha +1. Brevemente, neste estudo foram usados dados públicos de quatro segmentos de cana-de-açúcar (MATTIELLO et al., 2015) e os perfis de expressão dos genes foram agrupados com o software Mclust (SCRUCCA; FOP; RAFTERY, 2016). Os dados de Rossi (2022) derivam dos identificadores dos genes e dos arquivos de anotação de duas versões do genoma da variedade SP80-3280 em formato GFF (RIAÑO-PACHÓN; MATTIELLO, 2017; SOUZA et al., 2019). Uma tabela com a atribuição de cada um dos genes em cada um dos 99 “clusters” de co-expressão pode ser encontrada em https://doi.org/10.6084/m9.figshare.21748790, e foi o ponto de partida deste estudo.

3.2 Extração da região promotora de genes co-expressos

Para a extração da região promotora de cada um dos genes nos 99 grupos de genes co-expressos durante o desenvolvimento da folha +1 foi desenvolvido um script na linguagem Python usando a biblioteca PyRanges (STOVNER; SÆTROM, 2020). Este script (FindPromoter.py) lê os dados de anotação em formato GFF3 e os dados das sequências em formato FASTA dos dois rascunhos do genoma. Neste arquivo, se encontram as posições, ao longo dos contigs, dos éxons, sequências codificantes (CDS), regiões não traduzidas (UTR), entre outras. O script identifica a primeira CDS do gene, e extrai uma sequência de 1000 pares de bases (bp) a partir da primeira posição corrente acima (“upstream”) da CDS. Consideramos que as regiões de 1Kbp que sobrepõem com um outro gene não seriam válidas, e por isso, só foi mantida a região intergênica entre os dois genes, o gene de interesse e o gene “upstream” deste, o que poderia resultar em regiões promotoras menores que 1Kbp. Assim, neste estudo, foi definida como região promotora aquela iniciando no nucleotídeo imediatamente anterior ao códon de início do gene e até 1000bp “upstream”, até o limite do próximo gene “upstream”, ou até chegar ao limite do contig. O script para realizar a operação descrita está disponível em https://doi.org/10.6084/m9.figshare.21751520. Este script produz, então, um banco de dados em formato FASTA com todas as regiões promotoras dos genes do genoma, que usa como identificador da região promotora o mesmo identificador do gene a que

ele pertence. Um segundo script (LocateCluster.py, disponível em: https://doi.org/10.6084/m9.figshare.21751532), tem como entrada a lista de atribuição de genes em grupos de co-expressão gerado por Rossi (2022), e produz um arquivo para cada um dos grupos de co-expressão com os identificadores dos genes do grupo. Finalmente, com um terceiro script (ExtractSequenceFastafromList.py, disponível em: https://doi.org/10.6084/m9.figshare.21751538), é gerado um arquivo fasta, com as sequências promotoras de cada um dos genes, para cada um dos grupos de genes co-expressos.

3.3 Busca pelos motivos nos promotores dos genes co-expressos

A busca de motivos “de novo” foi realizada através do programa STREME (BAILEY, 2021). Esse programa utiliza como entrada um conjunto de sequências e gera uma saída com os motivos encontrados em forma de uma Position Weight Matriz (PWM). Foram realizadas duas buscas com esse programa para cada um dos 99 “clusters” identificados no estudo anterior (ROSSI, 2022), cuja sequências de entrada eram as regiões promotoras: uma busca usando como controle as mesmas sequências dos promotores, mas com a ordem das bases embaralhada de forma aleatória – “All”, e uma busca cujo controle eram as regiões promotoras de genes presentes em todos os outros “clusters” – “Exclusive”, esta segunda busca tinha como intuito facilitar a descoberta de motivos exclusivos de cada “cluster”. Para ambas as buscas o tamanho mínimo e máximo para o motivo foi de 6bp e 12bp, respectivamente, segundo estudos prévios em cana-de-açúcar (DINIZ et al., 2020; LIU et al., 2022). Os motivos identificados são representados por meio de uma PWM, utilizada para realizar comparações entre sequências posteriormente. A identificação dessas matrizes inclui o número do “cluster” onde foi achado, a estratégia de busca, um número serial e, ao final, a sequência consenso do motivo. Além desse dado, o programa também retorna as sequências que geraram aquele motivo.

Posteriormente, foi utilizado o programa TOMTOM (GUPTA et al., 2007) que realiza uma comparação dos motivos; essa ferramenta foi utilizada para identificar motivos similares e posteriormente agrupá-los. Isso foi realizado comparando todos os motivos resultantes do STREME contra eles mesmos, sendo os resultados de busca com *q-value* < 0,1 considerados significativos (*q-value* representa um ajuste no *p-value* utilizado para avaliar taxa de descoberta falsa). Em seguida, foi realizado um agrupamento com os motivos que encontraram correspondentes similares, baseando-

se numa estratégia de análise de redes com a criação de uma rede na qual os nós representavam os motivos, e as arestas representavam a similaridade entre pares de motivos. O software MCL (VAN DONGEN, 2000) foi utilizado com um valor de inflação de 21 (que foi o valor encontrado neste trabalho que maximiza o número de “clusters”), para identificar agrupamentos (“clusters”) na rede com alta similaridade. A rede foi visualizada com o software Gephi (BASTIAN; HEYMANN; JACOMY, 2009). Para cada “cluster” foi escolhido como motivo representante aquele com o maior valor de betweenness centrality e maior tamanho de sequência. Betweenness centrality é uma medida de centralidade dos nós da rede baseada no cálculo do caminho mais curto entre dois nós, seu valor representa a somatória das razões entre todos os caminhos mais curtos entre dois motivos qualquer que passam pelo motivo em que está sendo calculada a métrica (BRANDES, 2003). Os representantes de cada “cluster” foram selecionados para as próximas etapas de análise.

Com o intuito de identificar motivos previamente relatados na literatura, os motivos selecionados como representantes de cada “cluster” junto com aqueles que não formaram “clusters” (singletons), foram utilizados como entrada no programa TOMTOM, com o mesmo parâmetro de busca utilizado anteriormente, e comparados contra o banco de dados JASPAR (CASTRO-MONDRAGON et al., 2022), que reúne dados de sítios de ligação de FT. O resultado desta comparação também ajudou a identificar os FTs que possivelmente se ligariam ao motivo pesquisado, assim como a classe, família e identificador no banco de dados UniProt desses FTs. Dentre os motivos que tiveram correspondente no banco de dados, foram escolhidos aqueles que estavam presentes em pelo menos 30% das sequências de cada “cluster” de genes co-expressos, de acordo com os resultados do STREME, para as análises seguintes.

3.4 Análise dos FTs que, potencialmente, se ligam aos motivos selecionados

Para cada um dos motivos, foi considerada a família de FT mais representativa, ou seja, aquela com maior número de ocorrências identificada na comparação com o JASPAR. Em seguida, foi realizado uma pesquisa na literatura para os motivos que estavam representados em mais de 30% das sequências conforme a análise do STREME, utilizando-se informações obtidas pelo UniProt (BATEMAN et al., 2021), principalmente sua identificação no banco de dados Araport quando possível, para identificar as relações filogenéticas dos FTs já publicadas na literatura e suas funções.

3.5 Verificação dos motivos identificados: Escaneamento contra o genoma

A metodologia de escaneamento contra o genoma consistiu na realização de uma busca com motivos que tiveram “matches” significativos no banco de dados JASPAR, verificando sua ocorrência nos promotores de todos os genes anotados nos rascunhos do genoma da cana-de-açúcar da variedade SP80-3280, através da ferramenta FIMO (GRANT; BAILEY; NOBLE, 2011) utilizando *q-value* de 0,01. O FIMO, também integrante do “MEME suite”, avalia todas as sequências e procura por ocorrências dos motivos individualmente, porém se o número de correspondências encontradas atingir o valor máximo permitido, o FIMO descartará 50% das correspondências menos significativas, assim como as novas correspondências que estiverem abaixo do nível de significância (GRANT; BAILEY; NOBLE, 2011). O intuito desse escaneamento foi verificar a especificidade dos motivos encontrados dentro do genoma.

4 RESULTADOS

4.1 Tamanho da região promotora nos rascunhos do genoma da variedade SP80-3280

Nos dois rascunhos disponíveis do genoma da cana-de-açúcar da variedade SP80-3280 (RIAÑO-PACHÓN; MATTIELLO, 2017; SOUZA et al., 2019), existem 525.526 genes, dos quais podem ser extraídas as regiões promotoras com tamanho maior que zero pares de bases, sendo possível extrair uma região promotora de 1000 Kbp para 69,1% dos genes (Figura 2).

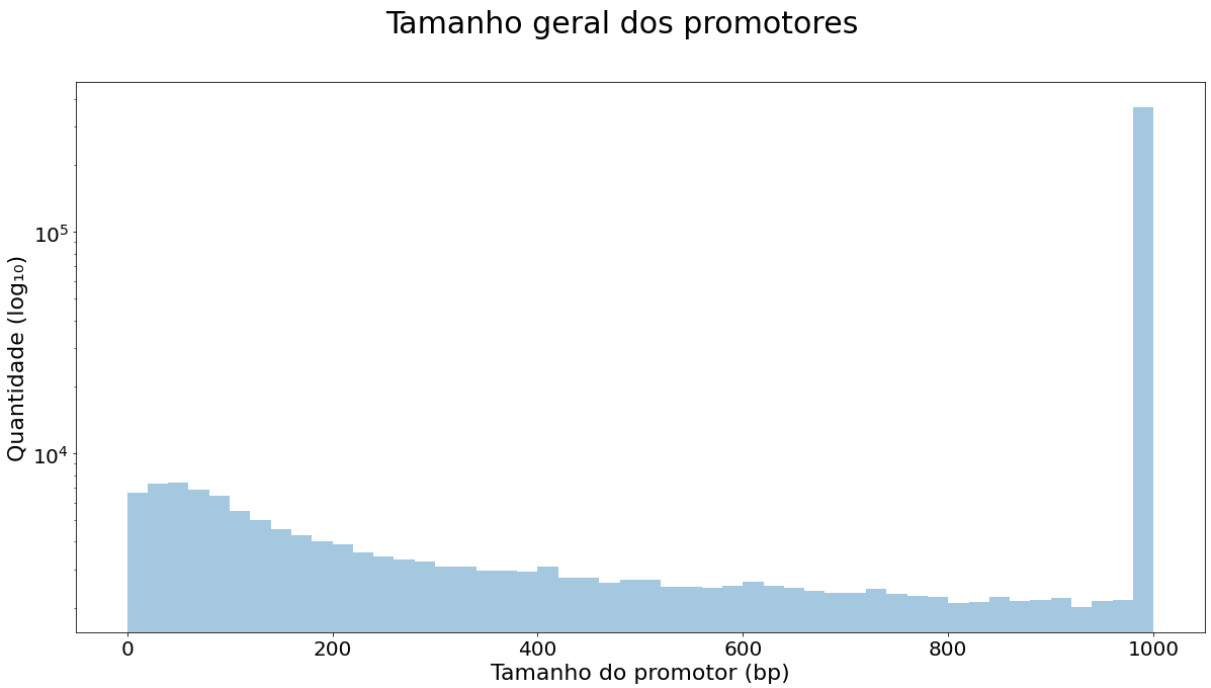

Figura 2. Histograma com os valores dos tamanhos dos promotores que puderam ser obtidos. Eixo y: quantidade de promotores com determinado valor em escala logarítmica (log_{10}); Eixo x: valor do tamanho dos promotores em pares de base.

Nota-se que a maior parte dos promotores apresentaram valores que permitiram a sequência das atividades.

4.2 Motivos identificados

Em relação aos resultados do STREME, para cada “cluster” de genes co-expressos houve em média 20,4 motivos encontrados, com um total de 2024 motivos, usando a estratégia de controle “All”, e uma média de 5,6, com um total de 555 motivos, usando a estratégia de controle “Exclusive”. Totalizando 2579 motivos encontrados. Os resultados do STREME são gerados em forma de uma PWM que

podem ser demonstradas como logotipo de sequência (Figura 3). Todas as PWM estão disponíveis em https://doi.org/10.6084/m9.figshare.21749246.

Figura 3. Exemplo de logotipo de sequência do motivo Cl_77_All_5-GARGRRGAGRGG obtido como resultado do STREME.

4.3 Agrupamento por similaridade dos motivos

Já quanto ao agrupamento dos motivos, o uso da ferramenta TOMTOM demonstrou que grande parte dos motivos encontrados apresentavam algum grau de similaridade com motivos dentro de um mesmo grupo ou de outros grupos de promotores de genes co-expressos. Apenas 307 motivos não encontraram um similar de acordo com os parâmetros selecionados, sendo 296 motivos resultantes da estratégia de busca "Exclusive" e 11 motivos da estratégia "All" na ferramenta STREME. Em contrapartida, houve alguns motivos que encontraram até 280 correspondências, como no caso do motivo “Cl_77_All_5-GARGRRGAGRGG”, cujo logotipo de sequência está na Figura 3.

As relações de similaridade entre os motivos foram representadas como uma rede, ou grafo, e visualizadas através da ferramenta Gephi (Figura 4). Nesta rede os nós representam os motivos, e seu diâmetro está representando o seu betweenness centrality. Nota-se que a maioria dos motivos apresentam pelo menos uma conexão com algum outro, fazendo deste um grafo altamente conectado. A partir dos dados utilizados para gerar essa rede, o MCL conseguiu agrupar os motivos em 298 “clusters”.

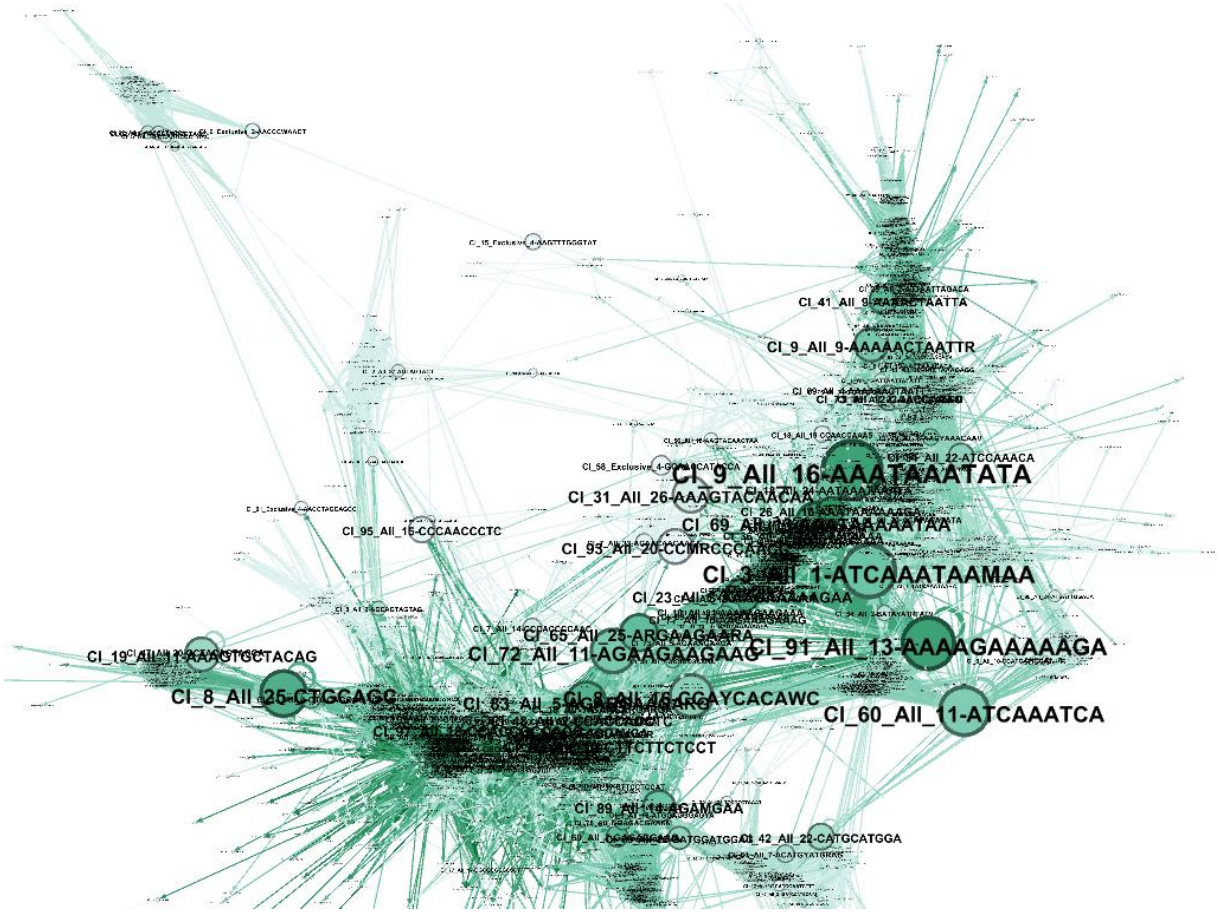

Figura 4. Grafo representando a similaridade entre os motivos desenvolvido através do software Gephi. Os nós representam os motivos e seu diâmetro está representando o seu betweenness centrality. Quanto mais escura a tonalidade do nó maior o seu grau, ou seja, maior o número de arestas conectadas a ele.

4.4 Comparação dos motivos contra banco de dados e análise

Dos 605 motivos selecionados, motivos únicos (307) somados aos motivos representativos dos clusters de motivos selecionados na etapa anterior (298), e comparados com 656 CREs de plantas anotados disponível no banco de dados JASPAR, através da ferramenta TOMTOM, apenas 98 apresentaram similaridade com algum dos perfis de ligação conhecidos de FT (arquivo correspondente disponível em https://doi.org/10.6084/m9.figshare.21749273). Dentre os resultados, foram observados CREs que apresentam sítio de ligação de 277 FTs. Os FTs estão classificados em 16 classes (Tabela 1) e distribuídos em 31 famílias (Tabela 2).

Tabela 2. Classificação dos FTs que apresentam sítio de ligação semelhante aos motivos descobertos e a quantidade de correspondências encontradas.

Classificação*	Quantidade de correspondências	Quantidade de motivos descobertos	Quantidade de motivos no JASPAR
AP2/EREBP	1912	89	102

AP2	1	1	6
ERF/DREB	1911	88	96
B3	17	7	28
ABI3	6	2	*4*
ARF	11	5	20
BBR/BPC	19	3	3
Basic helix-loop-helix factors (bHLH)	99	48	75
BES/BZR	8	5	6
TCP	12	7	30
Basic leucine zipper factors (bZIP)	63	38	47
Group A	8	8	9
Group B	1	1	1
Group C	4	2	2
Group D	15	9	11
Group G	6	5	5
Group H	1	1	2
Group K	2	1	1
Group S	25	10	10
C2H2 zinc finger factors	33	11	31
CAMTA	1	1	3
CH3	9	1	3
GRF	9	1	3
CPP	4	4	5
Fork head/winged helix factors	1	1	5
E2F	1	1	4
GCM domain factors	18	11	98
NAC	12	5	51
WRKY	4	4	45
WRKY-like_FRS/FRF	2	2	2
Heat shock factors	4	4	10
Homeo domain factors	18	15	30
HD-ZIP	11	9	25
PLINC	7	6	6
MADS box factors	33	18	21

M alpha	1	1	1
MIKC	32	17	20
Other C4 zinc finger-type factors	116	17	45
C4-GATA-related	2	2	11
DOF	64	12	31
LBD	50	3	3
Tryptophan cluster factors	62	24	141
GARP_G2-like	1	1	22
Myb	30	12	66
Myb-related	7	7	34
Trihelix	24	4	12

Obs.: Classe (negrito); Família (itálico).

Fonte: Autoria própria

Visando o aprofundamento nas análises de FTs, prosseguiu-se análise na literatura com 17 motivos que estavam com representação de mais de 30% em pelo menos um dos “clusters” que ele estava representando (Tabela 3). Cada motivo foi identificado junto com a família mais frequente dentre as famílias de FTs com sítio de ligação similar a ele, como observado na Tabela 4.

Tabela 3. Motivos representados por ID e suas respectivas sequência consenso e “clusters” no qual o motivo ou similar está incluso com representação de 30% ou maior entre as sequências das regiões promotoras.

ID do motivo representativo	“Cluster” de genes co-expressos cujo motivo identificado está presente em 30% ou mais regiões promotoras
Cl_11_All_2-CGTCGCCGCC	54
Cl_18_All_31-GRGGARGCGGCG	3
Cl_23_Exclusive_2-CCCATCATCTGC	9
Cl_26_All_16-AAATAAAAAAGA	72
Cl_26_All_8-GAWGAKGAWGAG	8 10 17 40
Cl_29_All_25-AAAAGAAGAA	29
Cl_35_All_10-GAGAGAGAAAG	35
Cl_52_All_2-CCKTCCCCGWC	52
Cl_53_Exclusive_2-GSGCSC	53
Cl_54_All_8-CCTCTCCCTCYC	54
Cl_63_All_22-AAATAAAAAAA	8 63 74
Cl_65_All_25-ARGAAGAARA	65
Cl_76_All_21-CTGCTGCTCC	93
Cl_87_All_5-CGCCGCCGCCGC	7 53 80 85 35
Cl_89_All_20-CCRCCACCACC	72
Cl_89_All_3-CGSCGSCGSCG	83
Cl_91_All_13-AAAAGAAAAAGA	2 4 5 7 11 13 16 21 26 29 30 35 36 40 41 43 47 49

	53
	58
	60
	67
	72
	77
	78
	80
	92
	96

Fonte: Autoria própria

Tabela 4. Motivos representados por ID e suas respectivas famílias de FT mais abundante dentre as correspondências com o banco de dados JASPAR.

ID do motivo	Família de FT obtida pelo JASPAR	Família de FT após correção	Subclassificações*	Referências
Cl_11_All_2-CGTCGCCGCC	ERF/DREB	AP2/ERF**	A-2 A-4 A-5 A-6 **B-1** B-2 **B-3** B-4 B-5 B-6	Sakuma et al. (2002)
Cl_18_All_31-GRGGARGCGGCG	ERF/DREB	AP2/ERF**	A-2 A-4 A-5 A-6 B-1 B-2 **B-3** B-4 B-5 B-6	Sakuma et al. (2002)
Cl_23_Exclusive_2-CCCATCATCTGC	MIKC	MADS-Box**	MIKC[c]	Theißen E GRAMZOW (2016)
Cl_26_All_16-AAATAAAAAAGA	DOF	DOF	Bb Cc Dd	Lijavetzky, Carbonero e Vicente-Carbajosa (2003)
Cl_26_All_8-GAWGAKGAWGAG	MIKC Myb-related	MADS-Box** Myb-related	MIKC[c] R-R-type	Theißen E GRAMZOW (2016) Yanhui et al. (2006)
Cl_29_All_25-AAAAGAAGAA	DOF	DOF	Aa Bb Cc Dd	Lijavetzky, Carbonero e Vicente-Carbajosa (2003)
Cl_35_All_10-GAGAGAGAAAG	BBR/BPC	BBR/BPC	Group I Group II	Theune et al. (2019)
Cl_52_All_2-CCKTCCCCGWC	ERF/DREB	AP2/ERF**	-	Sakuma et al. (2002)
Cl_53_Exclusive_2-GSGCSC	TCP	TCP	PCF (Classe I)	Martín-Trillo e Cubas (2010)
Cl_54_All_8-CCTCTCCCTCYC	BBR/BPC	BBR/BPC	Group I Group II	Theune et al. (2019)
Cl_63_All_22-AAATAAAAAAA	DOF	DOF	Aa Bb Dd	Lijavetzky, Carbonero e Vicente-Carbajosa (2003)
Cl_65_All_25-ARGAAGAARA	DOF	DOF	-	Lijavetzky, Carbonero e Vicente-Carbajosa (2003)
Cl_76_All_21-CTGCTGCTCC	GRF	GRF	A	Choi, Kim e Kende (2004)
Cl_87_All_5-CGCCGCCGCCGC	ERF/DREB	AP2/ERF**	A-2 A-4 A-5 A-6 B-1 B-2 **B-3** B-4 B-5 B-6	Sakuma et al. (2002)

Cl_89_All_20-CCRCCACCACC	ERF/DREB	AP2/ERF**	A-2 A-4 **A-5** A-6 **B-1** B-2 B-3 B-4 B-5 B-6	Sakuma et al. (2002)
Cl_89_All_3-CGSCGSCGSCG	ERF/DREB	AP2/ERF**	A-2 A-4 A-5 A-6 B-1 B-2 **B-3** B-4 B-5 B-6	Sakuma et al. (2002)
Cl_91_All_13-AAAAGAAAAAGA	DOF	DOF	Aa Bb Cc Dd	Lijavetzky, Carbonero e Vicente-Carbajosa (2003)

*Subclassificações em destaque obtiveram frequência absoluta de 9 ou maior para o motivo analisado.

**Identificação da família corrigida de acordo com a literatura escolhida.

Fonte: Autoria própria

Nota-se que alguns motivos estão presentes em mais de um cluster, em especial aqueles com correspondências na família DOF e AP2/ERF. Outros motivos não puderam ter a família de FTs que o reconhece identificadas pela falta dessa informação no banco de dados JASPAR.

Por fim, ressalta-se que houve a necessidade de correção nas identificações das famílias da Tabela 4, pois observou-se a ocorrência de identificações que diferiam da literatura utilizada para discussão.

4.5 Verificação dos motivos identificados: escaneamento contra o genoma

Em relação ao escaneamento geral do genoma, foram utilizados os 17 motivos descritos anteriormente como entrada no programa FIMO. Quando escaneados contra todos os promotores de SP80-3280, excetuando-se pelo motivo Cl_53_Exclusive_2-GSGCSC, que não encontrou nenhuma correspondência, os motivos não apresentavam especificidade evidente dentro dos “clusters” (https://doi.org/10.6084/m9.figshare.21749324).

5 DISCUSSÃO

5.1 Distribuição de tamanho dos promotores e similaridade de motivos encontrados

Foi possível adquirir promotores suficientes para o tamanho de 1000 pares de bases. Ainda assim, existe uma fração importante de genes, em que a região promotora foi menor que 1000bp. Essa variação pode ser devido à baixa qualidade dos dois rascunhos do genoma desta variedade, rascunhos que ainda estão muito fragmentados nas quais as montagens estão com valor de N50 muito baixo e quantidade de contigs alta. Além disso, existem genes com regiões intergênicas menores do que 1000bp (RIAÑO-PACHÓN; MATTIELLO, 2017; SOUZA et al., 2019).

5.2 Análise de similaridade dos motivos encontrados

No que concerne aos resultados da busca dos motivos nos promotores de genes co-expressos, observa-se que a estratégia “All” obteve um número médio de motivos maior que a estratégia “Exclusive”, o que está de acordo com o intuito inicial de filtrar para obtenção de motivos mais específicos. O programa utilizado STREME, ao final das análises, seleciona o motivo que melhor discrimina as sequências de entrada das sequências controle (BAILEY, 2021). Assim, motivos com sequências descobertas nas sequências de entrada similares àquelas presentes no controle são menos prováveis de serem recuperados. Dessa forma, na estratégia de busca que utiliza como controle as sequências promotoras de outros genes, ou seja, “Exclusive“, elementos recorrentes comuns entre promotores apresentam menos chances de serem detectados. Entretanto, na estratégia que utiliza o embaralhamento de forma aleatória das bases, ou seja, “All”, os elementos regulatórios presentes na região promotora apresentam suas sequências embaralhadas, portanto são passíveis de serem detectados. Justificando, assim, a necessidade de realizar comparações dos motivos por similaridade através da ferramenta TOMTOM, estratégia também adotada por Dillman et al. (2015), e, posteriormente, agrupamentos. Dentro desses agrupamentos, apesar de haver motivos presentes na Tabela 4 que compartilham alguns pares de bases na sequência consenso, o logotipo de sequência que os representa é diferente (Figura 5), o que resulta em diferentes correspondências de sítios de ligação de FT ao comparar com os bancos de dados.

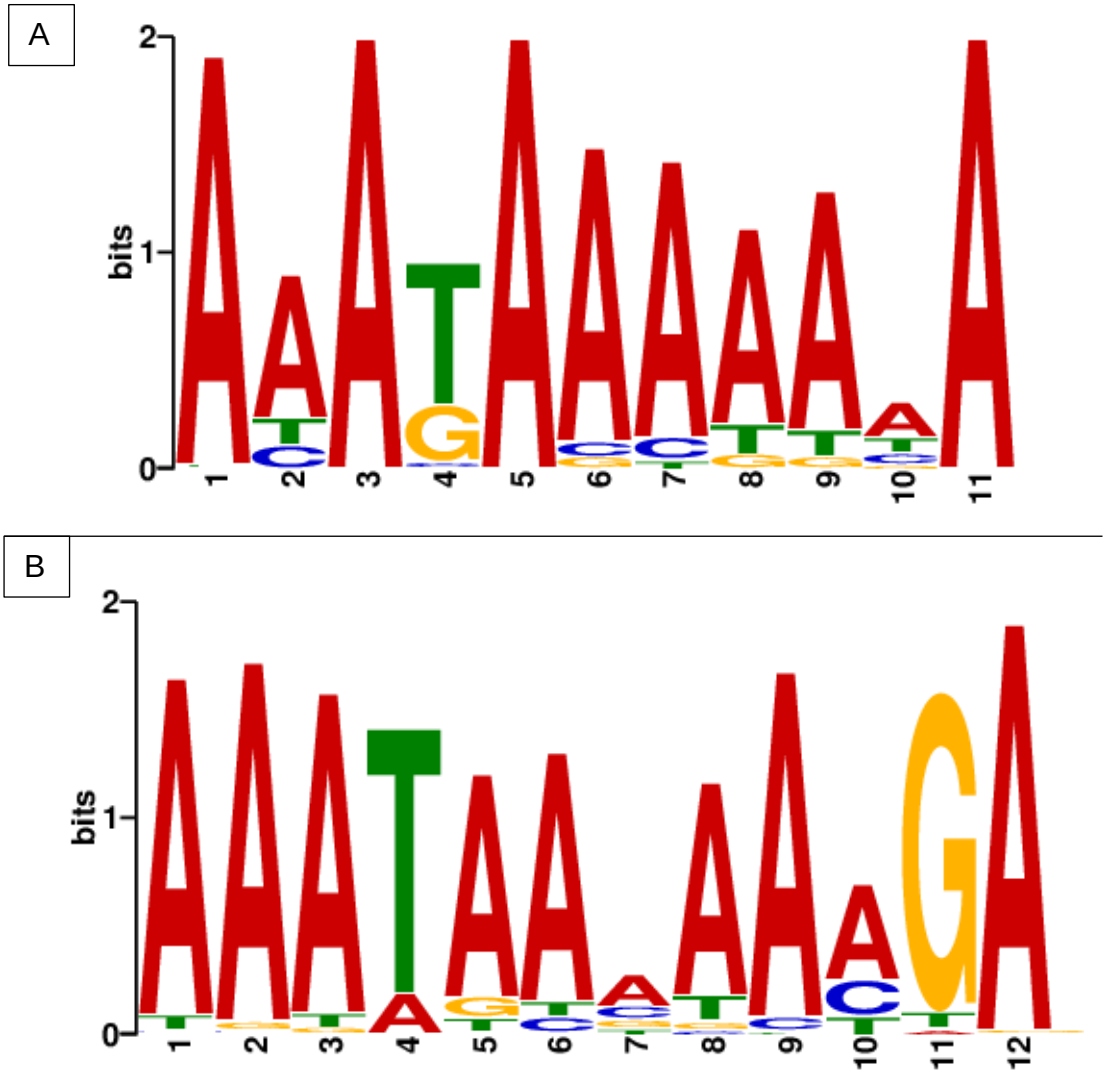

Figura 5. Comparação do logotipo de sequência dos motivos Cl_63_All_22-AAATAAAAAAA (A) e Cl_26_All_16-AAATAAAAAAGA (B).

5.3 Comparação dos motivos contra banco de dados

Nos resultados da comparação dos motivos contra o banco de dados, observa-se que menos de 20% dos motivos (98) utilizados na busca obtiveram correspondências, ou seja, encontraram similares dentre as sequências presentes no JASPAR. Dentre esses 98 motivos, apenas 14 foram encontrados através da estratégia “Exclusive”. Essa baixa correspondência pode ser decorrente da presença de promotores específicos para cada tecido (BIŁAS et al., 2016), visto que os dados derivam de análises específicas na folha da cana-de-açúcar, e da composição do banco de dados utilizado. A base de dados do banco de dados JASPAR é majoritariamente composta por sítios de ligação de FTs presentes em *Arabidopsis thaliana,* (86,5%), uma planta eudicotiledônea, em contrapartida há poucos dados sobre monocotiledôneas, à exemplo do milho que compõe apenas 6,3% dos dados

(CASTRO-MONDRAGON et al., 2022). Além disso, há ainda muitas sequências de elemento cis a serem analisadas e descobertas, o que traz a necessidade de novas pesquisas (BIŁAS et al., 2016), visto que algumas famílias apresentam baixa representação no banco de dados.

Apesar da baixa correspondência para os motivos inseridos na busca, pode-se observar que para um mesmo motivo é possível que ele seja reconhecido por FTs de famílias diferentes. Esse fato relaciona-se com a recorrência de sequências que são reconhecidas por famílias diferentes, a exemplo do motivo “Caixa-G”, cuja sequência é reconhecida por FTs das famílias bHLH e bZIP (EZER et al., 2017).

5.4 Principais famílias de FTs envolvidas na regulação do desenvolvimento da folha em cana-de-açúcar

Necessitou-se de uma adequação na análise das famílias, pois, comparando-se com os dados presentes em Gonzalez (2016b), o JASPAR classifica algumas famílias como classe e algumas subfamílias como famílias. Dentre os 17 motivos que foram utilizados no estudo e análise de literatura, as famílias de FTs mais presentes nas comparações com os dados do JASPAR foram: AP2/ERF, MADS-Box, DOF, Myb-related, BBR/BPC, TCP, GRF. A maioria dessas famílias também foram observadas sendo sobrerepresentadas em algumas partes da folha (MATTIELLO et al., 2015), como Myb e DOF.

Dos dois motivos presentes na Tabela 3 que não foi possível identificar as subclassificações da família do FT cujo sítio de ligação é similar a eles, o motivo Cl_52_All_2-CCKTCCCCGWC pode ser identificado como sendo reconhecido pelo FT Zm00001d031796 (TU et al., 2020), mas não há muitos dados na literatura sobre a sua classificação, porém é descrito por Ke et al. (2022) uma relação com a rota metabólica do metabolismo do açúcar e do amido, o que está em conformidade com Mattiello et al. (2015) que verificou enriquecimento de sequências na mesma rota metabólica. Já o outro motivo Cl_65_All_25-ARGAAGAARA, apesar de ter o FT correspondente classificado em um clado específico em análise por Chen e Cao (2015), não foi encontrada literatura relacionada a sua função.

5.4.1 AP2/ERF

A família AP2/ERF é uma das maiores famílias de fatores de transcrição encontradas em plantas (YAMASAKI, 2016) e apresenta alta quantidade de sítios de

ligação para essa família presente no banco de dados JASPAR. Assim, já era esperado que alguns motivos detectassem o sítio de ligação dos FTs dessa família. Conforme Feng et al. (2020), esses fatores de transcrição estão envolvidos em diversas funções: crescimento, desenvolvimento, regulação hormonal, controle da floração, senescência, destino do meristema da espigueta, iniciação da raiz, desenvolvimento e tamanho da folha, desenvolvimento de sementes/frutos e amadurecimento.

Essa família pode ser subdividida em 3 subfamílias baseados na estrutura geral: AP2, RAV e ERF. Sakuma et al. (2002) em seu estudo divide ERF em duas outras subfamílias ERF e CBF/DREB, gerando subclassificações dentro dessas: A1-A6 para ERF e B1-B6 para CBF/DREB. Já Nakano et al. (2006) estabelecem uma nova classificação filogenética, sendo a subfamília ERF dividida em 10 grupos, indicados de I à X.

No presente estudo, os subgrupos considerados sobrerepresentados para ERF foram o B-1 em Cl_11_All_2-CGTCGCCGCC e Cl_89_All_20-CCRCCACCACC; B-3 em Cl_11_All_2-CGTCGCCGCC, Cl_18_All_31-GRGGARGCGGCG, Cl_87_All_5-CGCCGCCGCCGC, Cl_89_All_3-CGSCGSCGSCG; e A-5 em Cl_89_All_20-CCRCCACCACC. A Tabela 5 apresenta a correspondência dos grupos para diferentes autores. Cabe ressaltar que, segundo Yamasaki (2016), os FTs DREB e ERF reconhecem sequências similares com diferenças sutis, assim a presença conjunta dessas duas subfamílias já era esperada para um mesmo motivo.

Tabela 5. Relações estre as subclassificações da família AP2/ERF para diferentes autores detectadas com frequência absoluta de 9 ou maior.

Subfamília[1]	Subgrupo[1]	Subgrupo[2]
DREB	A-5	II
ERF	B-1	VIII
ERF	B-3	IX e X

[1] Sakuma et al. (2002)

[2] Nakano et al. (2006).

Fonte: Autoria própria

O subgrupo A-5 apresenta FTs induzidos por estresse reportados em várias espécies (MIZOI; SHINOZAKI; YAMAGUCHI-SHINOZAKI, 2012). Esse grupo também contém FTs associados com resposta ao estresse abiótico atuando como repressores em genes associados à tolerância à seca e ao frio (MIZOI; SHINOZAKI; YAMAGUCHI-

SHINOZAKI, 2012). Assim, espera-se que os “clusters” enriquecidos com o sítio de ligação ao grupo A-5, estejam relacionados principalmente com estresse abiótico.

Os grupos B-1 e B-3 foram reportados sendo induzidos por estresse osmótico não-letal, sugerindo sua importância na adaptação a condições de estresse abiótico (MIZOI; SHINOZAKI; YAMAGUCHI-SHINOZAKI, 2012). Esses subgrupos também já foram reportados conjuntamente associados com resposta a patógenos (PRÉ et al., 2008). O subgrupo B3 é reportado como um dos principais agentes na regulação de vias de resistência a doenças (GUTTERSON; REUBER, 2004), biossíntese de alcaloides (YAMADA et al., 2020), na repressão do caminho de sinalização de ABA (PANDEY et al., 2005) e na resposta a diferentes estresses bióticos (GU et al. 2000). Em específico para B1, FTs pertencentes ao subgrupo VIII quando sobre-expressos induziram ou colaboraram na indução de morfologia da morte celular em *Arabidopsis* (OGATA et al., 2013). Dessa forma, pode haver uma relação entre repostas ao estresse abiótico e biótico e os “clusters” enriquecidos com sítios de ligação a FTs do grupo B-1 e B-3. Essa resposta tanto em relação ao estresse biótico quanto abiótico também foi verificada nas folhas de cana-de-açúcar por Mattiello et al. (2015).

5.4.2 DOF

DOF é uma família de fatores de transcrição que contém sítio de ligação em promotores de diversos genes, intervindo em vários aspectos do metabolismo de resposta às entradas ambientais para controlar as respostas de crescimento da planta (NOGUERO et al., 2013). Entretanto supõe-se que a maior parte dos sítios não sejam funcionais *in vivo* (YANAGISAWA, 2016) e que seu funcionamento, provavelmente, seja associado com outros fatores de transcrição. Essa combinação geraria a necessidade de especificidade do promotor (NOGUERO et al., 2013). Sendo assim, a presença de vários motivos que tenham similaridade com o sítio de ligação dessa família era esperada.

Lijavetzky, Carbonero e Vicente-Carbajosa (2003) propõem a divisão dos DOFs em 4 grupos: Aa, Bb, Cc, e Dd, utilizados nas identificações de DOFs deste trabalho, sendo todos os grupos encontrados nas correspondências na somatória dos motivos. Os motivos Cl_91_All_13-AAAAGAAAAAGA, Cl_63_All_22-AAATAAAAAAA e Cl_29_All_25-AAAAGAAGAA obtiveram correspondência com FTs do grupo Aa. Dentre os FTs com função conhecida nessa família, há o At1g07640 (AtDOF1.1) que está envolvido na biossíntese de um metabolito com função de defesa (SKIRYCZ et

al., 2006). Os grupos Bb, e Dd se encontraram presentes nas correspondências de todos os motivos cuja família mais representativa foi o DOF (Cl_26_All_16-AAATAAAAAAGA, Cl_63_All_22-AAATAAAAAAA e Cl_91_All_13-AAAAGAAAAAGA e Cl_29_All_25-AAAAGAAGAA). Dentre o grupo Bb há alguns FTs com função conhecida: AT5G02460 (AtDof5.1) e AT3G55370 (AtDof3.6), os quais apresentam papéis distintos. AtDof5.1 regula a polaridade adaxial-abaxial (KIM et al., 2010), já AtDof3.6 está associado com resposta à luz, pois é relatado modulando a sinalização de fitocromo e criptocromo em *Arabidopsis* (WARD et al., 2005), sendo que sua sobre-expressão resulta em defeitos de crescimento (Kang et al., 2003). Em relação ao grupo Dd, esse também se trata de um subgrupo cujos FTs apresentam funções diversas, dentre eles: AT1G69570 (CDF5), AT1G29160 (AtDof1.5), AT3G50410 (AtDof3.4) e AT5G66940 (AtDof5.8). CDF5 e AtDOf1.5 estão relacionados com floração fotoperiódica, o primeiro também atua na repressão da expressão de *CONSTANS* (FORNARA et al., 2009) e o segundo é sugerido ter o papel de regulador negativo da via de sinalização do fitocromo (PARK et al., 2003). Isso evidencia a atuação das forças evolutivas sobre os domínios não relacionados com a ligação ao DNA, como descrito por Gonzalez (2016a). Já AtDof3.4 atua no controle do ciclo celular e AtDof5.8 na diferenciação e desenvolvimento vascular (YANAGISAWA, 2016). Assim, observa-se que promotores com os motivos encontrados com similaridade aos sítios de ligação à família DOF estejam associadas a diversas funções importantes para o metabolismo da folha.

5.4.3 MADS-box

A família MADS-box, composta por FTs que contém o domínio MADS, é encontrada em quase todos os eucariotos (THEIßEN; GRAMZOW, 2016). Ressalta-se, porém, que o número de genes que codificam FTs MADS-box aumentou consideravelmente durante a evolução das plantas terrestres. Dentre suas funções pode-se citar desenvolvimento dos óvulos, sementes, flores e frutos. Essa família é subdividida em MADS-box tipo I e tipo II ou MIKC (THEIßEN; GRAMZOW, 2016), este último foi muito importante, pois suas duplicações propiciaram inovações evolucionárias. MIKC também é subdividido em $MIKC^C$, um grupo mais antigo que foi recuperado através das buscas dos motivos Cl_26_All_8-GAWGAKGAWGAG e Cl_23_Exclusive_2-CCCATCATCTGC, e MIKC* (HENSCHEL et al., 2002). Entre os

FTs que compõe MIKC^c, há aqueles envolvidos no desenvolvimento e na mudança do estágio vegetativo para o reprodutivo (THEIßEN; MELZER, 2016).

Entretanto, as funções de FTs pertencentes a esse grupo ainda estão sendo estudadas. Em uma pesquisa recente de Latif et al. (2022) observou-se que a sobre-expressão do gene AT5G62165 (AGL42), que foi identificado no presente estudo, atrasou a senescência foliar pela regulação negativa de genes NAC causadores de senescência em algodão. Portanto, os motivos similares ao seu sítio de ligação podem estar associados, também, com funções na folha.

5.4.4 BBR/BPC

A família Barley B-Recombinant/Basic Pentacysteine (BBR/BPC) é uma família recente na evolução das plantas (NAGATA; HOSAKA-SASAKI; KIKUCHI, 2016), tendo preferência para se ligar à motivos GAGA-DNA (MEISTER et al., 2004). Conforme Theune et al. (2019), essa família pode ser dividida em dois clados I e II, separados recentemente, pois esses clados são restritos a espermatófitas. Nesse estudo foram identificados três FTs pertencentes à família BBR/BPC em ambos os motivos Cl_35_All_10-GAGAGAGAAAG e Cl_54_All_8-CCTCTCCCTCYC: AT5G42520 (BPC6), AT2G01930 (BPC1) e AT4G38910 (BPC5), sendo apenas BPC1 pertencente ao clado I. Apesar de os motivos serem compostos de pares bases distintos, foi validada a ligação de BPC1 e BPC6 a sítios-alvo genômicos e sua afinidade para repetições estendidas dos nucleotídeos GA/TC *in vivo* (THEUNE et al., 2019).

Em relação às funções desses FTs, Theune et al. (2019) cita que BPC6 apresenta relação com a via sinalização de brassinoesteroides. Monfared et al., (2011), descreve sobre sobreposição de função dos FTs integrantes do clado I e II. Conforme o autor, plantas mutadas para BPC1, BPC2, BPC4 e BPC6 apresentaram defeitos vegetativos e reprodutivos, concluindo que os quatro FTs apresentam papel na senescência de inflorescência, crescimento laminar da lâmina foliar, tamanho do pecíolo e da lâmina foliar, especificação da forma da célula adaxial e abaxial e padrão estomático. Assim, observa-se nessa família um importante papel na determinação da morfologia da folha e pode-se inferir que os motivos com sítios de ligação similares à essa família também atuem no controle dessa função.

5.4.5 TCP

A família Teosinte branched1/Cincinnata/proliferating (TCP) faz parte de um grupo de FTs específicos de plantas (GONZÁLEZ-GRANDÍO; CUBAS, 2016). Essa família apresenta influência nos padrões de crescimento dos tecidos e órgãos durante o desenvolvimento da planta, especialmente nas folhas e flores (NICOLAS; CUBAS, 2016). Segundo Nicolas e Cubas (2016), esse controle de desenvolvimento da folha descrito para a família TCP está especificamente relacionado com controle do tempo de maturação e diferenciação. Os FTs TCP podem ser divididos em duas classes baseada nas diferenças do domínio TCP: classe I e classe II (MARTÍN-TRILLO; CUBAS, 2010). A classe I foi amplificada recentemente nos grupos vegetais, mas ainda não há análises filogenéticas suficientes para determinar seu grau de conservação entre as linhagens de plantas (GONZÁLEZ-GRANDÍO; CUBAS, 2016). Apenas um dos dois FTs recuperados pelo motivo Cl_53_Exclusive_2-GSGCSC pertencente à família TCP pôde ser identificado no banco de dados Araport, o FT AT1G35560 (TCP23) pertencente a classe I. Estudos indicam que os FTs dessa classe atuam redundantemente na regulação do crescimento, formato e maturação da folha. Mais especificamente em relação ao FT TCP23, Nicolas e Cubas (2016) citam sua atuação em folhas regulando o comprimento, largura, perímetro e área da lâmina, crescimento do pecíolo e diferenciação celular. Já Martín-Trillo e Cubas (2010) relatam que foi previsto a possibilidade de sua atuação no controle da transcrição dos genes do cloroplasto.

Assim, genes cujos promotores são apresentam sítios de ligação para essa família podem apresentar uma forte relação com o desenvolvimento da folha e dessa forma o motivo encontrado apresenta um forte vínculo com a origem dos dados de Mattiello et al. (2015).

5.4.6 GRF

A família GROWTH-REGULATING FACTOR (GRF) apresenta FTs relacionados com a manutenção de meristemas e promoção de proliferação celular em diversos órgãos em plantas (BAZIN et al., 2013), sendo detectados em tecidos com grandes padrões de expressão genes específicos para mitose. CHOI, KIM e KENDE (2004) em seu estudo agruparam os FTs da família GRF em três subfamílias: A, B e D. Os autores relatam que a subfamília A apresenta FTs que desempenham funções na semente e no estágio de muda em plantas. O FT OsGRF4 único

recuperado pela correspondência do motivo Cl_76_All_21-CTGCTGCTCC faz parte dessa subfamília.

Como já previsto para a subfamília A, estudos relatam a função de OsGRF4 no controle e desenvolvimento de grãos e no tamanho da panícula em arroz (DUAN et al., 2015; SUN et al., 2016). Entretanto, Li et al. (2018) em sua pesquisa observou que OsGRF4 confere regulação do crescimento, metabolismo de C e N em arroz. Além disso, os resultados do mesmo autor constataram influência do FT na assimilação de biomassa, folha e largura do caule, folha e colmo. Logo, apesar de o FT OsGRF4 ser encontrado associado, principalmente, com o desenvolvimento da semente, estudos recentes demonstram atuação no desenvolvimento e rotas metabólicas na folha, papel que pode ser sugerido aos genes cujo promotor apresenta sítios de ligação com esse FT.

5.4.7 Myb-related

A família MYB-related é descrita como parte da superfamília MYB. A superfamília MYB é composta por FTs que compartilham o domínio MYB, sendo que membros da família MYB-related são classificados por possuírem apenas uma ou duas repetições desse domínio (HONG, 2016). Dentre as funções conhecidas para os FTs dessa família há o controle da morfogênese celular, desenvolvimento do cloroplasto, associação com o ciclo circadiano (DU et al., 2013).

Essa família de FTs pode ser dividida em cinco subfamílias (YANHUI et al., 2006), dentre elas a R-R-type da qual um dos FTs encontrados na correspondência com o motivo Cl_26_All_8-GAWGAKGAWGAG faz parte. Conforme Yanhui et al. (2006), alguns FTs dessa subfamília estariam envolvidos com resposta de defesa em plantas, devido a verificação de alta expressão de genes de R-R-type associada níveis altos de ácido salicílico. Du et al. (2013) em seu artigo traz algumas funções do subgrupo R-R-type relacionadas ao ciclo circadiano, resposta ao estresse salino e promoção de senescência foliar, funções que possivelmente os genes com promotores contendo esse motivo também podem apresentar.

5.5 Verificação dos motivos identificados: escaneamento contra o genoma

O escaneamento do genoma através da ferramenta FIMO não gerou dados em relação à especialidade dos motivos à “clusters” específicos. Visto que a ferramenta FIMO não mostra todas as ocorrências de um motivo, não é possível observar as

quantidades totais para estimativa de relação entre motivos dentro de cada “cluster”. Além disso, a ferramenta FIMO realiza buscas por motivos individualmente e a co-expressão dos genes dentro de um conjunto pode estar atrelada a mais de um motivo (BIŁAS, 2016).

Já em relação ao motivo Cl_53_Exclusive_2-GSGCSC sua falta identificação no escaneamento pode estar atrelada ao fato de que a sequência é uma sequência curta e apresenta pouca informação funcional nas posições localizadas meio da sequência.

6 CONCLUSÃO

Conclui-se que a estratégia de avaliação de motivos em genes co-expressos foi eficaz na avaliação de CREs em cana-de-açúcar, visto que se pôde recuperar sítios de ligação de 98 motivos com correspondência no banco de dados de motivos anotados JASPAR, apesar de que não houve comprovação de especificidade de expressão a “clusters” determinados. Os motivos encontrados podem estar relacionados com funções no desenvolvimento, morfologia e resposta ao estresse de acordo com a subfamília. Além disso, são necessários mais estudos em monocotiledôneas, mais especificamente dos clados próximos à cana-de-açúcar, a fim de possibilitar um maior número de correspondências nos acervos de sítios de ligação com fatores de transcrição.

REFERÊNCIAS

AERTS, S. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. **Current Topics in Developmental Biology**, v. 98, p. 121–145, Jan. 2012.

BAILEY, T.L. STREME: accurate and versatile sequence *motif* discovery. **Bioinformatics**, v. 37, n. 18, p. 2834–2840, Sept. 2021.

BAILEY, T.L. et al. The MEME Suite. **Nucleic Acids Research**, v. 43, n. W1, p. W39–W49, 2015.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. **Gephi :** an open source software for exploring and manipulating networks visualization and exploration of large graphs. In: INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA. **Proceedings...** 2009.

BATEMAN, A. et al. UniProt: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, v. 49, n. D1, p. D480–D489, Jan. 2021.

BAZIN, J. et al. miR396 affects mycorrhization and root meristem activity in the legume M edicago truncatula. **The Plant Journal**, v. 74, n. 6, p. 920-934, 2013.

BIŁAS, R. et al. Cis-regulatory elements used to control gene expression in plants. **Plant Cell, Tissue and Organ Culture.** Amsterdam: Springer, 2016.

BOUCHER, C.; BROWN, D.G.; DUROCHER, S. On the structure of small *motif* recognition instances. In: INTERNATIONAL SYMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL. Berlin; Heidelberg: Springer, 2008. p. 269-281.

BRANDES, U. A faster algorithm for betweenness centrality. **Journal of Mathematical Sociology**, v. 25, n. 2, p. 163–177, 2003.

CASTRO-MONDRAGON, J.A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. **Nucleic Acids Research**, v. 50, n. D1, p. D165–D173, Jan. 2022.

CHOI, D.; KIM, J.H.; KENDE, H. **Whole genome analysis of the OsGRF gene family encoding plant-specific putative transcription activators in rice (*Oryza sativa* L.) Plant Cell Physiol**. [s.l: s.n.]. Disponível em: <https://academic.oup.com/pcp/article/45/7/897/1816319>.

COMPANHIA NACIONAL DE ABASTECIMENTO. **Nova estimativa de cana-de açúcar traz produção de 572,9 milhões de toneladas**. 2022. Disponível em: https://www.conab.gov.br/ultimas-noticias/4725-nova-estimativa-de-cana-de-acucar-traz-producao-de-572-9-milhoes-toneladas. Acesso em: 15 out. 2022.

CRICK, F.H. On protein synthesis. In: SYMP SOC EXP BIOL. 1958. p. 8.

DILLMAN, A.R. et al. Comparative genomics of Steinernema reveals deeply conserved gene regulatory networks. **Genome Biology**, v. 16, n. 1, Sept. 2015.

DINIZ, A.L. et al. Amino acid and carbohydrate metabolism are coordinated to maintain energetic balance during drought in sugarcane. **International Journal of Molecular Sciences**, v. 21, n. 23, p. 1–27, Dec. 2020.

DU, H. et al. Genome-wide identification and evolutionary and expression analyses of MYB-related genes in land plants. **DNA Research**, v. 20, n. 5, p. 437–448, Oct. 2013.

DUAN, P. et al. Regulation of OsGRF4 by OsmiR396 controls grain size and yield in rice. **Nature Plants**, v. 1, Dec. 2015.

EMBRAPA. **Brazilian science develops first non-GM gene-edited sugarcane of the world**. 2021. Disponível em: <https://www.embrapa.br/en/busca-de-noticias/-/noticia/66969890/brazilian-science-develops-first-non-gm-gene-edited-sugarcane-of-the-world#:~:text=Innovation%20Plant%20production-,Brazilian%20science%20develops%20first%20non%2DGM%20gene,edited%20sugarcane%20of%20the%20world&text=photo%2C%20Flex%20II-,Brazilian%20research%20has%20developed%20the%20sugarcane%20varieties%20Flex%20I%20and,of%20sucrose%20in%20plant%20tissues>. Acesso em: 18 dez. 2022.

EMPRESA DE PESQUISA ENERGÉTICA. **Balanço energético nacional 2021**: ano base 2020. Rio de Janeiro, 2021.

EZER, D. et al. The G-box transcriptional regulatory code in arabidopsis. **Plant Physiology**, v. 175, n. 2, p. 628–640, Oct. 2017.

FENG, K. et al. **Advances in AP2/ERF super-family transcription factors in plant**. **Critical Reviews in Biotechnology.** Taylor and Francis, 2020.

FORNARA, F. et al. *Arabidopsis* DOF transcription factors act redundantly to reduce CONSTANS expression and are essential for a photoperiodic flowering response. **Developmental Cell**, v. 17, n. 1, p. 75-86, 2009.

GAO, L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. **Nature Genetics**, v. 51, n. 6, p. 1044–1051, June 2019.

GARSMEUR, O. et al. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. **Nature Communications**, v. 9, n. 1, Dec. 2018.

GONZALEZ, D. H. Introduction to transcription factor structure and function. In: **Plant transcription factors:** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016a. p. 3–11.

GONZALEZ, D.H. **Plant transcription factors :** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016b.

GONZÁLEZ-GRANDÍO, E.; CUBAS, P. TCP transcription factors: evolution, structure, and biochemical function. In: **Plant transcription factors:** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016. p. 139–151.

GRANT, C. E.; BAILEY, T. L.; NOBLE, W. S. FIMO: Scanning for occurrences of a given *motif*. **Bioinformatics**, v. 27, n. 7, p. 1017–1018, 2011.

GU, Yong-Qiang et al. Tomato transcription factors Pti4, Pti5, and Pti6 activate defense responses when expressed in *Arabidopsis*. **The Plant Cell**, v. 14, n. 4, p. 817-831, 2002.

GUPTA, S. et al. Quantifying similarity between *motifs*. **Genome Biology**, v. 8, n. 2, Feb. 2007.

GUTTERSON, N.; REUBER, T.L. Regulation of disease resistance pathways by AP2/ERF transcription factors. **Current Opinion in Plant Biology**, Aug. 2004.

HENSCHEL, K. et al. Two ancient classes of MIKC-type MADS-box genes are present in the moss *Physcomitrella patens*. **Molecular Biology and Evolution**, v. 19, n. 6, p. 801-814, 2002.

HONG, J. C. General aspects of plant transcription factor families. In: **Plant** transcription factors: evolutionary, structural and functional aspects. [s.l.] Elsevier Inc., 2016. p. 35–56.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção Agropecuária**. 2022. Disponível em: https://www.ibge.gov.br/explica/producao-agropecuaria/. Acesso em: 10 out. 2022.

JUVEN-GERSHON, T.; KADONAGA, J. T. **Regulation of gene expression via the core promoter and the basal transcriptional machinery**. **Developmental Biology.** Academic Press, Mar. 2010.

KANG, Hong-Gu et al. Target genes for OBP3, a Dof transcription factor, include novel basic helix-loop-helix domain proteins inducible by salicylic acid. **The Plant Journal**, v. 35, n. 3, p. 362-372, 2003.

KE, J. et al. Combinatorial analysis of transcription and metabolism reveals the regulatory network associated with antioxidant substances in waxy corn. **Food Quality and Safety**, Jan. 2022.

KIM, Hyung-Sae et al. The DOF transcription factor Dof5. 1 influences leaf axial patterning by promoting Revoluta transcription in *Arabidopsis*. **The Plant Journal**, v. 64, n. 3, p. 524-535, 2010.

KORNBERG, R.D. The molecular basis of eukaryotic transcription. **Proceedings of the National Academy of Sciences**, Washington, v. 104, n. 32, p. 12955-12961, Aug. 2007.

LATIF, A. et al. Overexpression of the AGL42 gene in cotton delayed leaf senescence through downregulation of NAC transcription factors. **Scientific Reports**, v. 12, n. 1, p. 21093, Dec. 2022.

LI, S. et al. Modulating plant growth–metabolism coordination for sustainable agriculture. **Nature**, v. 560, n. 7720, p. 595–600, Aug. 2018.

LIJAVETZKY, D.; CARBONERO, P.; VICENTE-CARBAJOSA, J. Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families. **BMC Evolutionary Biology**. 2003.

MARTÍN-TRILLO, M.; CUBAS, P. TCP genes: a family snapshot ten years later. **Trends in Plant Science**, Jan. 2010.

MATTIELLO, L. et al. Physiological and transcriptional analyses of developmental stages along sugarcane leaf. **BMC Plant Biology**, v. 15, n. 1, Dec. 2015.

MEISTER, R.J. et al. Definition and interactions of a positive regulatory element of the *Arabidopsis* INNER NO OUTER promoter. **The Plant Journal**, v. 37, n. 3, p. 426-438, 2004.

MIZOI, J.; SHINOZAKI, K.; YAMAGUCHI-SHINOZAKI, K. AP2/ERF family transcription factors in plant abiotic stress responses. **Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, Feb. 2012.

MONFARED, M.M. et al. Overlapping and antagonistic activities of BASIC PENTACYSTEINE genes affect a range of developmental processes in *Arabidopsis*. **Plant Journal**, v. 66, n. 6, p. 1020–1031, 2011.

MISHRA, P. et al. Identification of cis-regulatory elements associated with salinity and drought stress tolerance in rice from co-expressed gene interaction networks. **Bioinformation**, v. 14, n. 03, p. 123-131, Mar. 2018. Biomedical Informatics.

NAGATA, T.; HOSAKA-SASAKI, A.; KIKUCHI, S. The evolutionary diversification of genes that encode transcription factor proteins in plants. In: **Plant transcription factors:** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016. p. 73–97.

NAKANO, T. et al. Genome-wide analysis of the ERF gene family in arabidopsis and rice. **Plant Physiology**, v. 140, n. 2, p. 411–432, 2006.

NICOLAS, M.; CUBAS, P. The role of TCP transcription factors in shaping flower structure, leaf morphology, and plant architecture. In: **Plant transcription factors:** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016. p. 249–267.

NISSANI, N.; ULITSKY, I. Unique features of transcription termination and initiation at closely spaced tandem human genes. **Molecular Systems Biology**, v. 18, n. 4, Apr. 2022.

NOGUERO, M. et al. The role of the DNA-binding One Zinc Finger (DOF) transcription factor family in plants. **Plant Science**, Aug. 2013.

OGATA, T. et al. Analysis of the cell death-inducing ability of the ethylene response factors in group VIII of the AP2/ERF family. **Plant Science**, v. 209, p. 12–23, Aug. 2013.

PANDEY, G.K. et al. ABR1, an APETALA2-domain transcription factor that functions as a repressor of ABA response in. *Arabidopsis* **Plant Physiology**, v. 139, n. 3, p. 1185-1193, 2005.

PARK, Don Ha et al. The *Arabidopsis* COG1 gene encodes a Dof domain transcription factor and negatively regulates phytochrome signaling. **The Plant Journal**, v. 34, n. 2, p. 161-171, 2003.

PRÉ, M. et al. The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense. **Plant Physiology**, v. 147, n. 3, p. 1347-1357, 2008.

RIAÑO-PACHÓN, D.M.; MATTIELLO, L. Draft genome sequencing of the sugarcane hybrid SP80-3280. **F1000Research**, v. 6, p. 861, June 2017.

ROSSI, V.S. **Identificação e análise de co-expressão in sílico de genes codificadores de pro-teínas associadas a transcrição e de enzimas associadas ao metabolismo de carboidratos em cana-de-açúcar (*Saccharum* spp.) da cultivar SP80-3280**. 2022. 133 p. Dissertação de Mestrado em Ciências – Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, 2022.

SAKUMA, Y. et al. DNA-binding specificity of the ERF/AP2 domain of *Arabidopsis* DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. **Biochemical and Biophysical Research Communications**, v. 290, n. 3, p. 998–1009, 2002.

SCRUCCA, L.; FOP, M.; RAFTERY, A.E. **mclust 5:** clustering, classification and density estimation using gaussian finite mixture models. Disponível em: <http://cran-logs.rstudio.com>. Acesso em: 14 dez. 2022.

SFORÇA, D.A. **Variação genética em poliploides complexos:** desvendando a dinâmica alélica em cana-de-açúcar. Campinas: Universidade Estadual de Campinas, 2019.

SKIRYCZ, A. et al. DOF transcription factor AtDof1. 1 (OBP2) is part of a regulatory network controlling glucosinolate biosynthesis in *Arabidopsis*. **The Plant Journal**, v. 47, n. 1, p. 10-24, 2006.

SOUZA, G.M. et al. Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. **GigaScience**, v. 8, n. 12, Dec. 2019.

STOVNER, E.B.; SÆTROM, P. PyRanges: efficient comparison of genomic intervals in Python. **Bioinformatics**, v. 36, n. 3, p. 918–919, Feb. 2020.

SUN, P. et al. OsGRF4 controls grain shape, panicle length and seed shattering in rice. **Journal of Integrative Plant Biology**, v. 58, n. 10, p. 836–847, Oct. 2016.

THEIßEN, G.; GRAMZOW, L. Structure and evolution of plant MADS domain transcription factors. In: **Plant transcription factors:** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016. p. 127–138.

THEUNE, M.L. et al. Phylogenetic analyses and GAGA-*motif* binding studies of BBR/BPC proteins lend to clues in GAGA-*MOTIF* recognition and a regulatory role in brassinosteroid signaling. **Frontiers in Plant Science**, v. 10, Apr. 2019.

TRUJILLO-MONTENEGRO, J.H. et al. Unraveling the genome of a high yielding colombian sugarcane hybrid. **Frontiers in Plant Science**, v. 12, Aug. 2021.

TU, X. et al. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. **Nature Communications**, v. 11, n. 1, Dec. 2020.

VAN DONGEN, S. A cluster algorithm for graphs. **Information Systems [INS]**, 2000.

WARD, J.M. et al. The Dof transcription factor OBP3 modulates phytochrome and cryptochrome signaling in *Arabidopsis*. **The Plant Cell**, v. 17, n. 2, p. 475-485, 2005.

WEINER, P. Linear pattern matching algorithms**.** In: ANNUAL SYMPOSIUM ON SWITCHING AND AUTOMATA THEORY, 14., 1973. IEEE. p. 1-11.

WITTKOPP, P.J.; KALAY, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. **Nature Reviews Genetics,** Jan. 2012.

YAMADA, Y. et al. Genome-wide identification of AP2/ERF transcription factor-encoding genes in California poppy (*Eschscholzia californica*) and their expression profiles in response to methyl jasmonate. **Scientific Reports**, v. 10, n. 1, Dec. 2020.

YAMASAKI, K. Structures, functions, and evolutionary histories of DNA-binding domains of plant-specific transcription factors. In: **Plant transcription factors:** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016. p. 57–72.

YANAGISAWA, S. Structure, function, and evolution of the dof transcription factor family. In: **Plant transcription factors:** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016. p. 183–197.

YANAGISAWA, S. The Dof family of plant transcription. **Trends in Plant Science**, v. 7, n. 12, p. 555-560, 2002. Disponível em: <http://plants.trends.com1360-1385/02/$-seefrontmatter>. Acesso em: 14 dez. 2022.

YANHUI, C. et al. The MYB transcription factor superfamily of *Arabidopsis*: Expression analysis and phylogenetic comparison with the rice MYB family. **Plant Molecular Biology**, v. 60, n. 1, p. 107–124, Jan. 2006.

ZAMBELLI, Federico; PESOLE, Graziano; PAVESI, Giulio. *Motif* discovery and transcription factor binding sites before and after the next-generation sequencing era. **Briefings in bioinformatics**, v. 14, n. 2, p. 225-237, 2013.

ZHENG, Y. et al. Volatile metabolomics and coexpression network analyses provide insight into the formation of the characteristic cultivar aroma of oolong tea (*Camellia sinensis*). **LWT**, v. 164, Jan. 2022.

BIBLIOGRAFIA CONSULTADA

GHORBANI, R. et al. Genome-wide analysis of AP2/ERF transcription factors family in Brassica napus. **Physiology and Molecular Biology of Plants**, v. 26, n. 7, p. 1463–1476, July 2020.

RODRIGUEZ, R.E. et al. Growth-regulating factors, a transcription factor family regulating more than just plant growth. In: **Plant transcription factors:** evolutionary, structural and functional aspects. Amsterdam: Elsevier, 2016. p. 269–280.

TIAN, Feng; YANG, De-Chang; MENG, Yu-Qi; JIN, Jinpu; GAO, Ge. PlantRegMap: charting functional regulatory maps in plants. **Nucleic Acids Research**, Oxford, v. 48, p. D1104-D1113, 8 nov. 2019. Oxford University Press (OUP).