**UNIVERSIDADE FEDERAL DE PERNAMBUCO**

**CENTRO DE INFORMÁTICA**

**RELATÓRIO DE DISCIPLINA**



**Alunos: Filipe Cordeiro de Medeiros Azevedo**

**Vinicius Emanuel**

**Recife,**

**2018**

**Sumário**

[Resumo 3](#_Toc512927877)

[Estado da Arte de Contextualização do Tema 4](#_Toc512927878)

[1. Conhecendo os genes, sua estrutura e expressão 4](#_Toc512927879)

[3. Ferramentas computacionais para análises de elementos cis-regulatórios em plantas 11](#_Toc512927880)

[4. O porquê de estudar TFs e seus alvos, ECRs 12](#_Toc512927881)

[5. Análises de enriquecimento e a busca por significado biológico em dados de Ômicas 14](#_Toc512927882)

[6. Objetivo geral 18](#_Toc512927883)

[7. Objetivos específicos 18](#_Toc512927884)

[8. Material e métodos 18](#_Toc512927885)

[9. Resultados 18](#_Toc512927886)

[10. Referências Bibliográficas 18](#_Toc512927887)

# **Resumo**

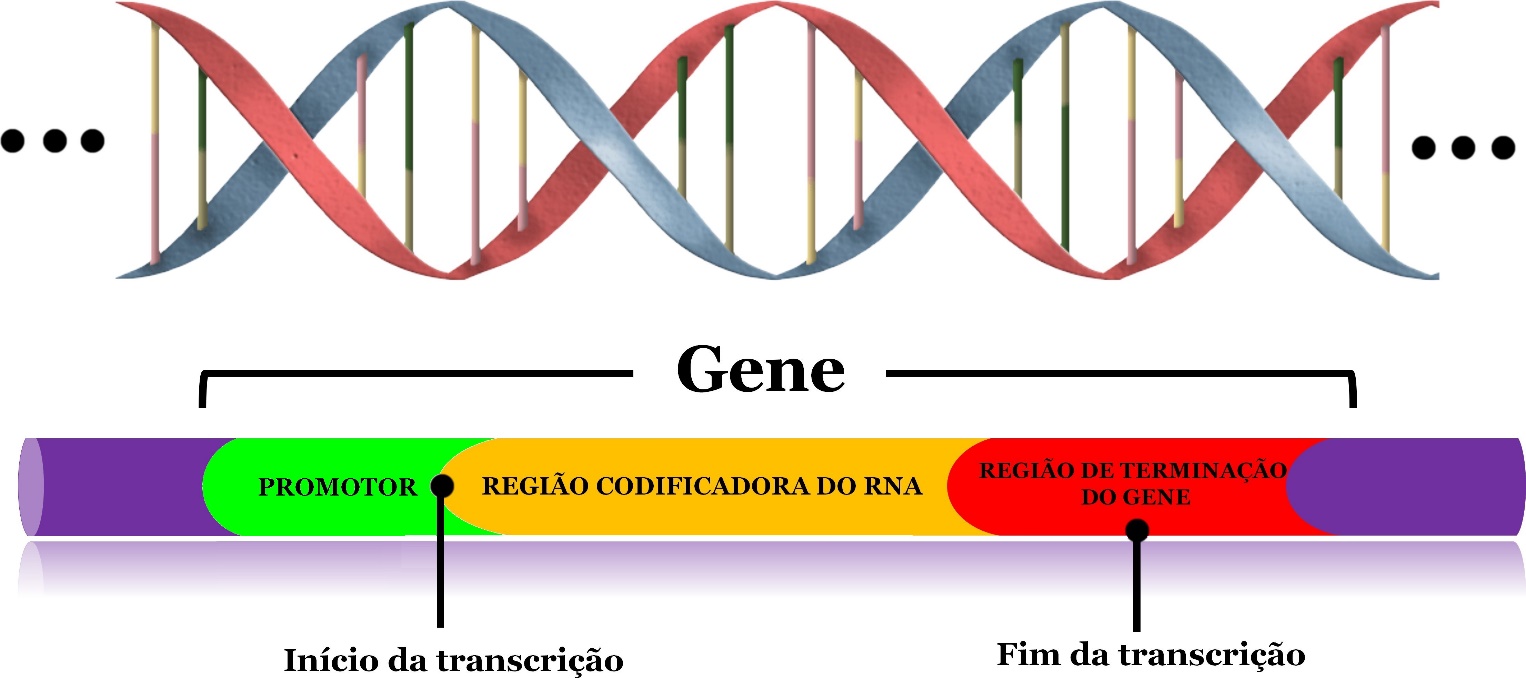
Genes são estruturas que contém a informação para a manifestação da vida. Sua expressão deve ser corretamente regulada para que os organismos se adaptem e enfrentem as oscilações do meio em que habitam. Tal processo regulatório ocorre, principalmente, numa região gênica denominada de promotor. Essa estrutura ancora sequencias curtas de nucleotídeos, chamada de elementos *cis*-regulatórios (ECRs), que são reconhecidas por proteínas denominadas fatores de transcrição (FTs), as quais são responsáveis pela expressão (ativação) dos genes. Para um determinado gene ser ativado por um dado FT é necessário que ele possua em seu promotor o elemento ECR reconhecido por essa proteína. Diferentes famílias de FTs reconhecem diferentes ECRs. Atualmente, existem diversos métodos moleculares para a mineração de associação entre ECR-FT. Tais descobertas levaram a construção de diversos bancos de dados que alocam matrizes de posição, as quais indicam os motivos de sequencias reconhecidos por diferentes famílias de FTs. Nesses bancos, o usuário fornece a sequência de nucleotídeos de um promotor e ferramentas computacionais tratam de processá-lo, visando encontrar os padrões pré-determinados nas matrizes. Tal ação atribui a um determinado gene, FTs pelos quais podem ser regulados, agregando dados sobre a sua dinâmica e informações sobre seu potencial biotecnológico. Os bancos de dados disponíveis, entretanto, oferecem um *output* pouco intuitivo, indicando somente os ECRs ancorados num dado promotor e, por conseguinte, quais os FTs envolvidos na regulação do mesmo, em formato tabular. O presente trabalho tem como objetivo a construção da plataforma Phytoprom, uma ferramenta online mineradora de ECRs de promotores de plantas, conectada via API com a base de dados online Phytozome, alocadora de genomas completos de 78 espécies de plantas. Tal ferramenta tem como diferencial a apresentação gráfica e intuitiva dos *outputs* gerados, bem como a aplicação pioneira de análises de enriquecimento em relação a essa temática. Esse método utiliza abordagens estatísticas para identificar grupos de elementos significativamente enriquecidos (super-representados) em relação a outros. Dessa forma, o uso da ferramenta PhytoProm possibilitará indicar quais grupos de FTs estão enriquecidos em relação a um conjunto de promotores, e por consequência genes, escrutinados. Esse resultado poderá ser utilizado para intuir a participação de genes, cuja função é desconhecida, na resposta a diferentes estresses bem como sugerir possíveis redes de corregulação.

# **Estado da Arte de Contextualização do Tema**

# **Conhecendo os genes, sua estrutura e expressão**

Genes são estruturas que contém a informação para a manifestação da vida. Coloquialmente, podem ser entendidos como ingredientes de uma receita que, combinados, resultam em diferentes pratos indicados para ocasiões distintas. Cientificamente, são designados como segmentos de uma molécula de DNA (ácido desoxirribonucleico), unidades de informação cujo conjunto forma um organismo funcional. Diferentes combinações de genes resultam em organismos com diferentes propriedades.

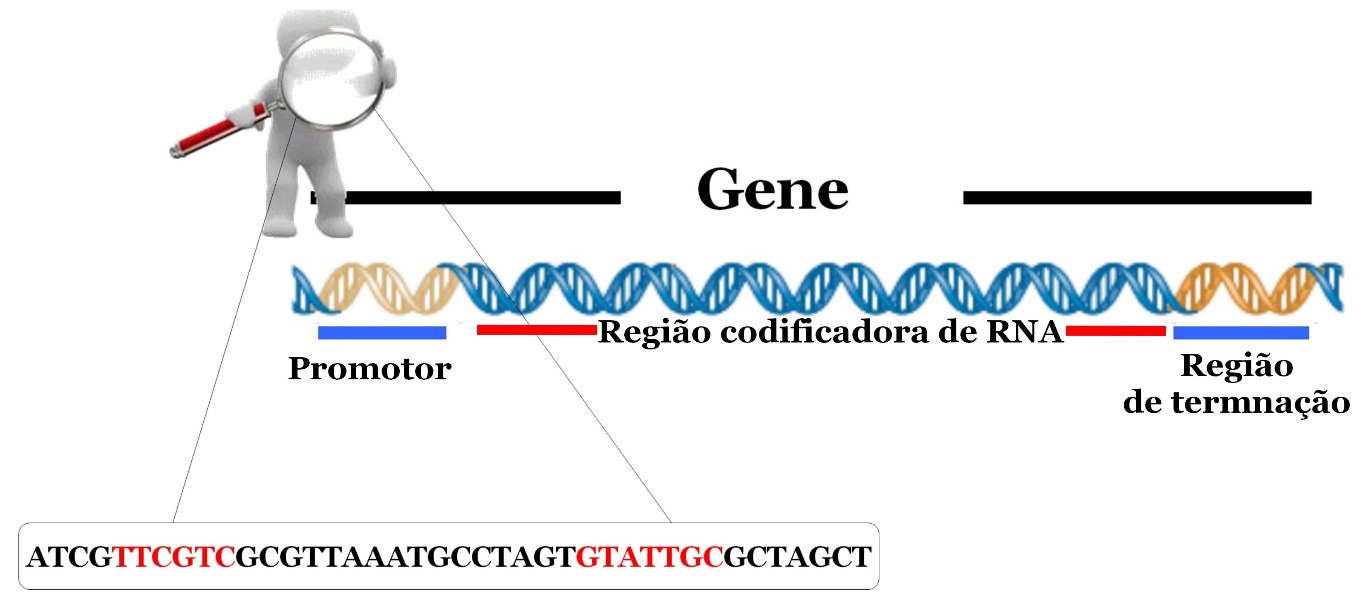
Didática e simploriamente, a anatomia dos genes – de eucariotos a procariotos – pode ser dividida em três seções fundamentais (**Figura 1**), que exercem funções diferenciadas na sua dinâmica:



**Figura 1.** Anatomia básica de um gene procarioto ou eucarioto, apresentando suas três partes básicas indicadas por cores distintas, bem como as regiões de iniciação e terminação de sua transcrição.

1. **Promotor**: região do DNA que indica o início de um gene. É um ponto fundamental do controle de sua expressão;
2. **Região codificadora de RNA**: ancora um conjunto de nucleotídeos que codifica a informação necessária para a formação de uma proteína;
3. **Região de terminação do gene**: ancora um conjunto de nucleotídeos que, quando “lidos” por uma RNA polimerase, fazem com que ela se dissocie do DNA, terminando o processo de transcrição do gene.

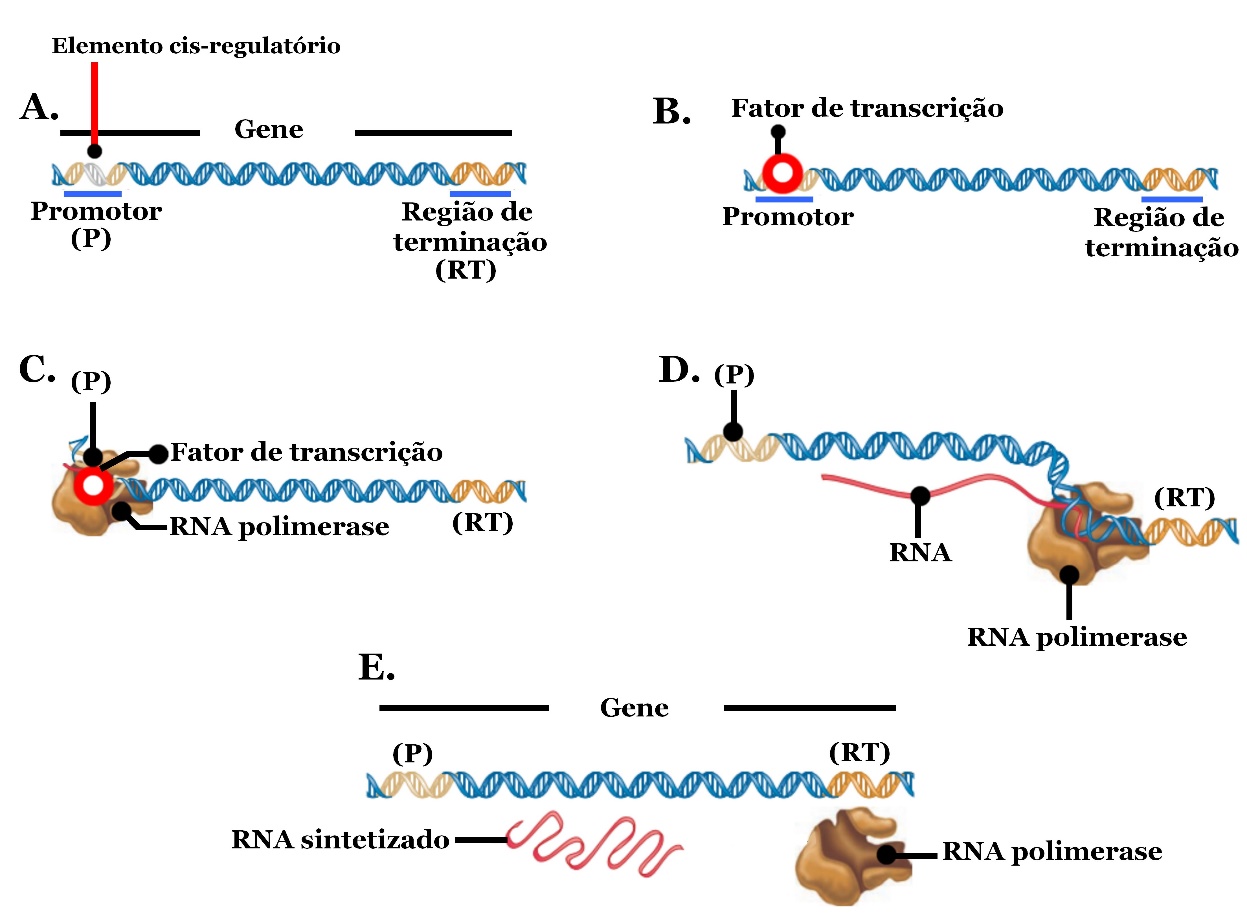
O promotor é considerado, essencialmente, uma região regulatória. Nele, se ligam proteínas denominadas Fatores de Transcrição (FTs). Tal ligação é executada por meio do reconhecimento de sequencias específicas de cinco a 20 nucleotídeos de extensão (Rombatus *et al*., 2003), chamadas de elementos *cis-*regulatórios(ECRs), situados na estrutura do promotor (**Figura 2**).



**Figura 2.** Apresentação da estrutura básica de um gene, enfatizando sua região promotora (retângulo) e destacando (em vermelho) os elementos *cis*-regulatórios hipotéticos que ela ancora.

Em organismos eucariotos (por exemplo: plantas, humanos, etc.), a presença de FTs associados a ECRs nos promotores é condição *sine qua non* para que um gene seja expresso (isto é, que seja capaz de produzir um RNA e sua respectiva proteína, caso possua tal informação). Nessa região gênica, FTs atuam posicionando adequadamente e ativando a enzima RNA-polimerase, responsável pela síntese de todos os tipos de RNAs (codificadores de proteínas ou não). Quando um determinado gene “A” necessita ser expresso, pois seu produto é requerido para uma dada reposta fisiológica, um FT, regulador do referido gene e ativado em resposta a condição fisiológica mencionada, se anexa a seu respectivo ECR no promotor do gene “A”. Em contrapartida, quando o produto do gene “A” não se faz mais necessário, o FT é desconectado do ECR por meio de interação com outras moléculas reguladoras. Esse mecanismo de liga/desliga é de suma importância na dinâmica energética de um organismo vivo, pois resulta em economia energética, tornando os sistemas vivos mais eficientes e adaptados aos seus ambientes. Fato similar ocorre em nossas residências, onde lâmpadas se encontram ligadas apenas em cômodos ou situações onde se fazem necessárias. Analogamente, para o nosso organismo é tão custoso deixar um gene - não requerido a todo momento – ligado permanentemente, quanto é para os nossos bolsos deixar as luzes da casa ligadas sem necessidade.

A **Figura 3** apresenta a dinâmica de anexação de um FT a seu respectivo ECR em um promotor, com posterior anexação da RNA polimerase e síntese de um RNA codificador de uma determinada proteína.



**Figura 3**. Síntese de RNA (também chamada de transcrição do DNA) de um gene e suas diferentes etapas. **A.** Apresentação da anatomia básica de um gene, destacando um elemento *cis*-regulatório (em cinza) no promotor; **B.** Reconhecimento do elemento *cis*-regulatório (ECR) por fator de transcrição (FT); **C.** Após anexação do FT ao promotor, o mesmo conecta uma RNA polimerase a essa região e a ativa; **D.** RNA polimerase sintetiza RNA, utilizando a região que codifica informação no DNA como molde; **E.** RNA polimerase chega à zona de terminação do gene, a reconhece e se dissocia.

Resumidamente, um determinado FT reconhece seu respectivo ECR num promotor, conectando-se a ele (**Figura 3A e 3B)**; posteriormente, é feito o reconhecimento e acoplamento do FT à RNA polimerase (**Figura 3C)**; em seguida, essa enzima é posicionada na região inicial de um gene e é ativada pelo FT (**Figura 3C)**. Após isso, o FT pode ser desconectado do promotor ou conectar outra RNA polimerase ao mesmo; o próximo passo é a execução da transcrição, onde a RNA polimerase irá transcrever a informação genética do DNA para uma molécula de RNA (**Figura 3D)**; após transcrever todo o conteúdo informacional do gene, a RNA polimerase chega a região chamada de “terminação”. Nessa região, há um conjunto de nucleotídeos que indica que o gene acaba naquele ponto, sendo que, dessa forma, a RNA polimerase é desconectada do DNA (**Figura 3E)**. Assim, chega ao fim o processo de transcrição de um determinado gene.

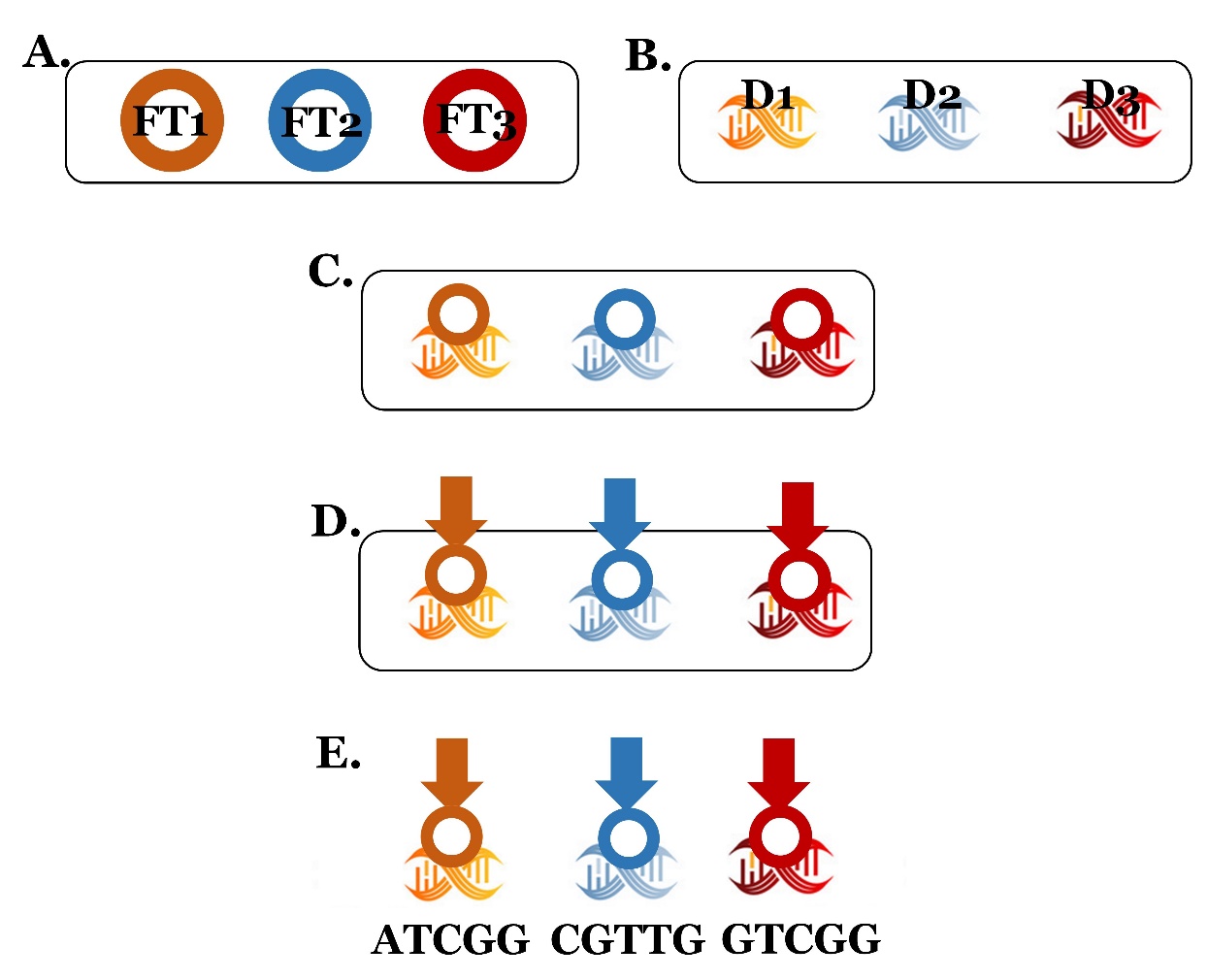
1. **Métodos de identificação da interação entre ECRs e FTs: do experimental ao computacional**

ECRs são sequencias de DNA ditas não codificantes (ou seja, não apresentam informação para síntese de uma proteína). Entretanto, tais elementos são de suma importância, pois são requeridos para a correta expressão de um gene (Li *et al*., 2015). Dessa forma, a identificação de ECRs é imperativa e de suma importância, pois agrega informação à dinâmica regulatória das unidades informacionais dos genomas. Em caso de genes não anotados (ou seja, sem identificação), isso pode dar indícios sobre suas potenciais funções, bem como indicar algum valor biotecnológico, dependendo do ECRs minerados e dos FTs que são associados aos mesmos.

Há, basicamente, duas formas de detecção da interação entre FTs e ECRs: metodologias baseadas em evidências experimentais; e aquelas baseadas em estratégias computacionais.

As experimentais subdividem-se em dois grupos:

* Métodos *in vivo*, tais como imunoprecipitação da cromatina seguida de sequenciamento de regiões alvos do DNA (ChIP-Seq) (**Figura 4**). Nessa metodologia (Xie *et al*., 2011), diferentes fatores de transcrição (**Figura 4A**) podem ser testados em relação aos seus ECRs. Para isso, o DNA genômico é fragmentado (**Figura 4B**), e colocado em uma solução com os FTs (**Figura 4C**). Uma vez que ambos estão em solução, os FTs se ligarão aos fragmentos de DNA que ancoram seus respectivos ECRs (**Figura 4C**). Após isso, são colocados na solução anticorpos específicos dos FTs sob análises (**Figura 4D**). Uma vez que tais anticorpos reconhecem seus alvos, os FTs associados aos seus respectivos fragmentos de DNA são precipitados (**Figura 4D**); por fim, os fragmentos de DNA são separados de seus FTs e sequenciados (**Figura 4E**).



**Figura 4.** Dinâmica das etapas da metodologia Chip-Seq para detecção da interação entre fatores de transcrição (FTs) e seus respectivos elementos *cis*-regulatórios (ECRs). **A.** Diferentes FTs cuja capacidade de ligação a diferentes ECRs será analisada; **B.** Fragmentos de DNA a serem analisados contra os FTs mencionados no item “A”. **C.** Ligação de FTs com fragmento de DNA contendo seu respectivo ECR alvo; **D.** Anticorpos (setas) reconhecem e se ligam a seus respectivos FTs e os precipitam da solução; **E.** Posteriormente, anticorpos e FTs são desconectados dos fragmentos de DNA analisados, os quais são sequenciados revelando o contexto de nucleotídeos que compõe o ECR associado a cada FT analisado.

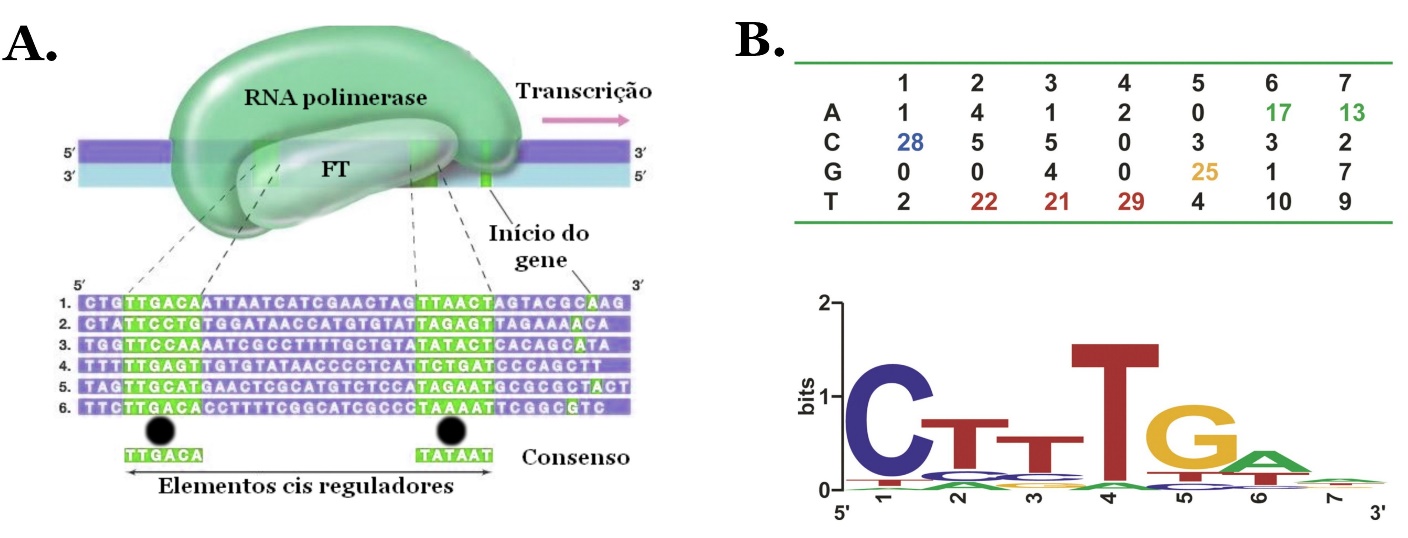
* Métodos *in vitro*, abrange estratégias baseadas na ligação de grandes conjuntos de fragmentos de DNA a proteínas de interesse, no caso, os FTs. Neste item, destacam-se as técnicas de Evolução Sistemática de Ligantes por Enriquecimento Exponencial (SELEX, do inglês “*Systematic Evolution of Ligands by Exponential Enrichment*”) e microarranjos de ligação a proteínas (PBM, do inglês *Protein-binding Microarrays*). Para informações mais aprofundadas sobre os três métodos aqui apresentados consultar Xie *et al*. (2011).

As estratégias computacionais empregam métodos *ab initio* (Trindade *et al*., 2005). Esses possuem em suas equações parâmetros empíricos ou semi-empíricos (ou seja, são derivados diretamente dos princípios teóricos, não incluindo dados experimentais). Tais abordagens visam encontrar padrões comuns de motivos de nucleotídeos (potenciais ECRs) em regiões promotoras de genes que apresentam perfil de expressão similar (Pilpel *et al*., 2001; Roth *et al*., 1998; Vilo *et al*., 2000; Vilo e Kivinen, 2001). Essa estratégia tem como ideia central o fato de que diferentes genes que apresentam um padrão de expressão similar, podem estar sob regulação de um mesmo grupo de FTs. Isso, portanto, pode ser evidenciado por meio de motivos de nucleotídeos similares existentes em seus promotores. Bussemaker *et al*. (2001) propuseram um método baseados nessa estratégia (correlação de valores de expressão gênica e ocorrências de motivos de nucleotídeos similares nos promotores).

Outras abordagens alternativas para o agrupamento de genes candidatos com perfis de expressão estão disponíveis. Essas consideram: grupos de genes com funções similares (Hughes *et al*. 2000; Zhu e Zhang 2000); genes com magnitude expressão similar sob certas condições (Jensen e Knudsen, 2000); agrupamento de diversos genes afetados em estudos de *knock-out* (Hughes *et al*. 2000); e genes envolvidos na mesma via metabólica (Brazma *et al*. 1998; van Helden *et al*. 1998).

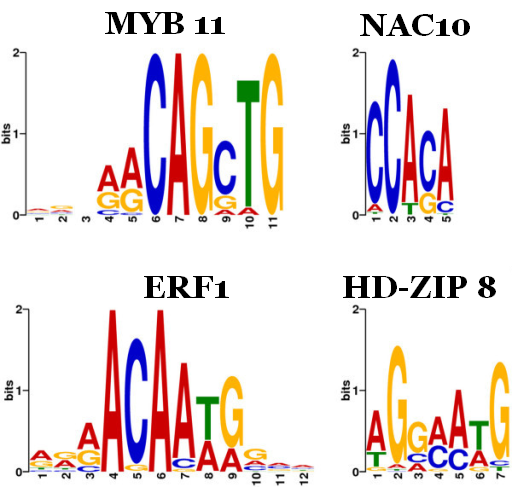
A associação ECR-FT (tanto experimental quanto computacional) fornece um modelo de ligação preferencial do FT ao DNA. Tal informação, por sua vez, pode ser usada para prever ECRs em outras sequências gênicas e, por consequência, prever por quais FTs determinados genes podem ser regulados. A representação computacional das preferências de ligação de FTs evoluiu ao longo dos anos, de simples sequências de consenso (**Figura 5A**) para as matrizes de frequência de posição (PFM, do inglês “*Position Frequency Matrices*”) (**Figura 5B**). Sequencias consenso são produzidas a partir da intersecção de nucleotídeos apresentados em um conjunto de sequencias (**Figura 5A**). Os nucleotídeos que apresentassem maior frequência numa dada posição seriam indicados na sequência consenso como nucleotídeo padrão para aquele respectivo ponto (**Figura 5A**). Ao fim da análise seria gerada uma sequência única e absoluta (**Figura 5A**).

Uma PFM, por sua vez, representa as sequencias de DNA as quais um TF se liga, por meio da apresentação do número de nucleotídeos reconhecido (extensão do ECR), bem como da contagem do número de vezes que um determinado nucleotídeo foi identificado em determinada posição (**Figura 5B,** partesuperior). Isso indica que não há sequência absoluta (consenso) (**Figura 5A**) de ECRs, mas sim um conjunto de nucleotídeos, de determinada extensão, que são reconhecidos por um dado FT. As referidas matrizes podem ser convertidas em matrizes de peso de posição (PWMs, do inglês “*Position Weight Matrices*”) (**Figura 5B**, parte inferior), também conhecidas como matrizes de pontuação específicas, que são modelos probabilísticos que podem ser usados para prever ECRs em sequências de DNA [para maiores informações sobre o processo consultar Stormo (2013)]. Em PWMs, (**Figura 5B**, parte inferior) quanto maior a “altura” do nucleotídeo representado numa posição, maior é a probabilidade do referido nucleotídeo ocupar a mesma e vice-versa.



**Figura 5.** Tipos de representação dos elementos cis-regulatórios (ECRs). **A.** metodologia antiga, apresentando os ECRs em formato de sequência consenso (ex. TTGACA e TATAAT); **B.** padrão atual de representação baseando-se em modelos probabilísticos. Parte superior apresenta uma matriz de frequência de posição; parte inferior apresenta matriz de peso de posição.

É válido ressaltar que diferentes FTs reconhecem diferentes ECRs, ou seja possuem diferentes matrizes de reconhecimento (**Figura 6**). Tal propriedade dos FTs é usada para sua categorização em diferentes famílias (Wingender *et al*., 2015).



**Figura 6.** Fatores de transcrição de diferentes famílias (MYB, NAC, ERF e HD-ZIP) acompanhados de seus respectivos elementos *cis*-regulatórios representados na forma de matriz de peso de posição (PWMs).

# **Ferramentas computacionais para análises de elementos cis-regulatórios em plantas**

Atualmente, existem diversas ferramentas de bioinformática disponíveis para identificar ECRs em potenciais promotores. Essas empregam, ou disponibilizam publicamente PWMs, oriundas de análises dos métodos apresentados no “item 2”. As ferramentas mais utilizadas para plantas estão descritas na **Tabela 1**.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tabela 1.**Ferramentas computacionais comumente utilizadas para mineração de elementos *cis*-reguladores em promotores de genes de plantas. | | | |
| **Ferramenta** | **Acesso** | **Site** | **Referência** |
| AGRIS | livre | http://agris-knowledgebase.org/AtcisDB/ | Yilmaz *et al*. (2011) |
| ATENA | livre | http://www.bioinformatics2.wsu.edu/Athena | ­- |
| ATHMAP | livre | http://www.athamap.de/ | Hehl *et al*. (2016) |
| PLACE | livre | http://www.dna.affrc.go.jp/htdocs/PLACE/ | Higo *et al*. (1999) |
| PlanPAN | livre | http://plantpan2.itps.ncku.edu.tw/ | Chow *et al*. (2015) |
| JASPAR | livre | http://jaspar.genereg.net/collection/core/ | Khan *et al*. (2018) |
| TRANSFAC | pago | http://genexplain.com/transfac/ | ­- |

TRANSFAC é um banco de dados comercial (pago) bastante abrangente e bem conhecido. Apresenta cerca de 6000 PWMs para diversos clados eucarióticos. PLACE e JASPAR foram criados para facilitar a detecção de ECRs. Nesses bancos online, sequencias de nucleotídeos de promotores, em formato fasta, podem ser escaneadas quando à existência dos referidos elementos. Entretanto, há limitações estratégicas que impedem a análise em larga escala. PlantPAN é outra importante base online de dados de ECRs, contando atualmente com 1143 PWMs de 10 diferentes espécies. Nessa base de dados são aceitas até 20 sequencias do tipo FASTA para análise. AGRIS, Athena e ATHMAP são bases de dados de mineração de ECRs na planta modelo Arabidopsis. São consideradas as mais úteis para tal vegetal.

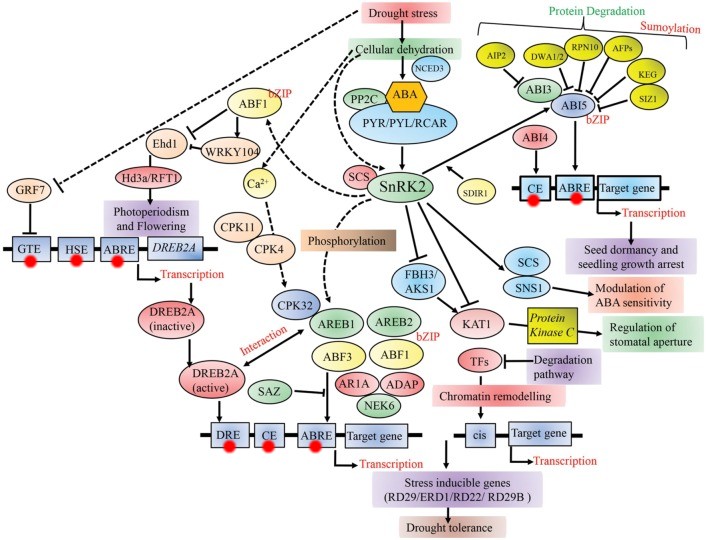
Os bancos acima mencionados têm como limitações o fato de serem pagos (TRANSFAC), ou não realizarem análises em alta escala (TRANSFAC, PLACE, PlantPAN, JASPAR, AGRIS, Athena e ATHMAP) ou serem espécie-específicos (AGRIS, Athena e ATHMAP, específicos de Arabidopsis). A estratégia padrão empregada para análises em alto rendimento, consiste em recuperar PWMs disponíveis e escrever códigos e executam análises em linha de comando (exemplo: Wei e Chen, 2018). Tal ação, é restrita a um grupo de pesquisadores com escopo científico específico (informatas e bioinformatas). Além disso, o output (tabular) disponibilizado pela maioria das ferramentas é útil, porém pouco intuitivo e apresentável em manuscritos científicos. Dessa forma, observa-se que é imperativo o desenvolvimento de novas ferramentas que suprimam essas limitações, realizando análises em alto rendimento, com boa estética na apresentação do resultado final, bem como abrangente, permitindo a análise de outras espécies ainda não estudadas a partir dessa temática.

# **O porquê de estudar TFs e seus alvos, ECRs**

Existem ao menos 60 famílias de fatores de transcrição em plantas (http://planttfdb.cbi.pku.edu.cn/). Em média, genomas vegetais contêm entre 1000 a 3000 genes codificadores de TFs, o que corresponde entre 5 a 15% de seu conteúdo informacional (Seo *et al*., 2015). Tais proteínas são responsáveis, primariamente, pela reorganização da fisiologia interna das células perante condições não favoráveis. São elas que ativam genes específicos cujos produtos modificam o ambiente celular, propiciando o enfrentamento da nova condição. Para que elas exerçam tal função, entretanto, é necessário a existência dos ECRs nos promotores dos genes. Por exemplo, caso um fator de transcrição reconheça um ECR de sequência “AATG”, ele só poderá atuar sobre genes que possuam essa sequência em seu promotor. Assim, uma vez que um determinado FT tenha sua participação documentada numa dada condição (ex: reposta à seca) no vegetal, genes que possuem o ECR reconhecido por esse FT, potencialmente, participam do processo de resposta a essa condição. A associação FTs|ECRs|genes cria uma rede de informação regulatória de suma importância para entender a dinâmica dos organismos, decodificando-os funcionalmente.

Estudos com mutantes, seja com perda de função, usando silenciamento gênico induzido por vírus (Hsieh *et al*., 2013), bem como estudos com transgênicos, superexpressando genes codificadores de TFs (Yao *et al*., 2016), são comumente adotados para investigar as funções dos TFs. A participação dessas famílias proteicas na reposta a estresses é bem sedimentada [para uma revisão consultar Hoang *et al*. (2017)]. Adicionalmente, análises de transcriptomas de plantas sob condições estressantes fornecem uma visão global da expressão de genes TF envolvidos nessas situações. Exemplo disso, pode ser observado em Song *et al*. (2016). Tais autores relataram que vários TFs reguladores do crescimento da raiz também foram induzidos em reposta a diferentes níveis de déficit hídrico, destacando o papel do crescimento da raiz como resposta adaptativa a períodos de seca.

Estudos abrangendo TFs têm levado ao rascunho de diversas vias de regulação dos mesmos. Essas apresentam quais TFs são ativados em situações específicas, bem como quais genes regulados por esses. Exemplo disso, pode ser visto na **Figura 7**, a qual apresenta a participação de diferentes famílias de TFs e, por conseguinte, ECRs na resposta à desidratação radicular.



**Figura 7.** Modelo esquemático da regulação de diferentes FTs sobre genes alvos os quais desempenham papéis-chave na reposta à desidratação radicular em plantas. Os FTs ativados AREB / ABF, AREB1, AREB2, ABF3 e ABF1 ligam-se à região promotora dos genes alvos e ativam a sua expressão em resposta ao stress de desidratação. GRF7 suprime a expressão de DREB2A, que é um fator chave na expressão gênica independente de ABA. Linhas tracejadas indicam possíveis ações. **Legenda:** PP2C-PYR/PYL/RCAR: *pyrabactin resistance1/PYR1-like/regulatory components of ABA receptors*; PP2C: *protein phosphatase 2C*; CE: *coupling element; GTE, GRF7-targeting cis-element*; HSE: *heat shock element*; FTs: fatores de transcrição. Bolas vermelhas indicam ECRs (elementos *cis*-reguladores). DRE: *dehydration-responsive element*; ABRE: *ABA-responsive element*; CE: *coupling elemento*; GTE: *GRF7-targeting cis- element*; HSE: *heat shock element;* Adaptado de Joshi *et al*. (2016).

# **Análises de enriquecimento e a busca por significado biológico em dados de Ômicas**

A possibilidade de sequenciar completamente genomas deu origem a uma subárea da genética: a Genômica (em inglês, *Genomics*). Segundo Kuska (1998), há indicações que esse termo foi criado por Thomas H. Roderick, para abranger o mapeamento, sequenciamento e análise da função de genomas inteiros. Para Yadav (2007) o termo inglês “OME” deriva da palavra em sânscrito “OM”, que significa plenitude ou totalidade. Atualmente, o sufixo inglês “OMIC” é colocado ao fim de epítetos de áreas já existentes nas ciências da vida, ou para tecnologias, indicando a realização de abordagens globais ou em alto escala/rendimento.

Historicamente, além da genômica, a transcriptômica (estudo de todos os transcritos produzidos em um dado momento, sob determinado estímulo e em um tecido específico) também tirou proveito das metodologias NGSs (*Next Generation Sequencing*). Esse termo descreve plataformas que produzem grandes quantidades (tipicamente, milhões) de curtas sequencias (25 a 400 pb) de DNA ou RNA, aumentando a produção dessas e reduzindo os custos em ordens de magnitude (Mardis, 2008). Adicionalmente, o desenvolvimento de tecnologias de proteômica, as quais estudam globalmente um conjunto de proteínas produzidas em um dado momento, sob determinado estímulo e em um tecido específico, completaram as análises globais dos componentes do Dogma Central da Biologia Molecular. Assim, Genômica, Transcriptômica e Proteômica são considerados as primeiras Ômicas.

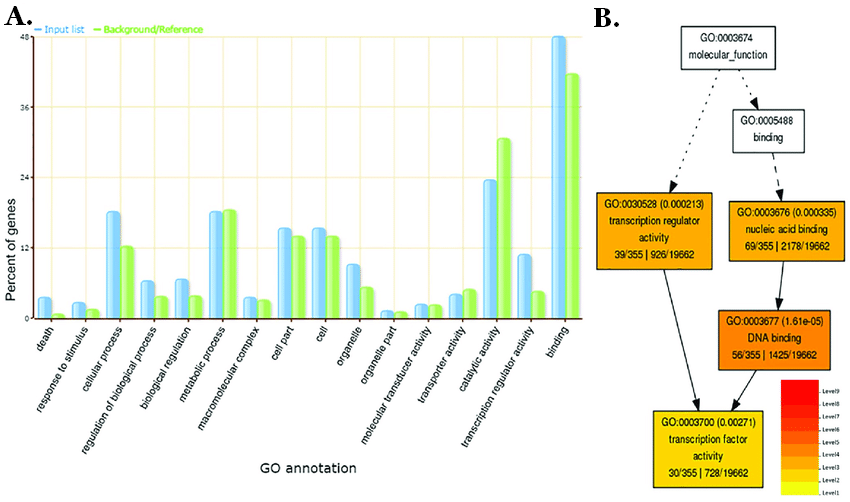
O estágio final de muitas análises em Ômicas é a produção de uma lista de biomoléculas (DNA ou RNA ou proteínas) de interesse. A estratégia de “Análise de Enriquecimento” se tornou uma abordagem de grande empregabilidade, devido a sua capacidade de fornecer *insights* valiosos sobre a efetiva participação de um grupo de elementos específicos (genes ou RNAs ou proteínas) em um dado processo fisiológico. O resultado de tal abordagem é alcançado por meio de várias etapas de análise (Tipney e Hunter, 2010). Num primeiro momento, executa-se a identificação sistemática das biomoléculas sob análise, bem como a descrição de suas potenciais funções. Após isso, compara-se a distribuição dos termos obtidos dentro do conjunto sob estudo com a distribuição de fundo desses termos. Entende-se “distribuição de fundo” (ou DF) como um valor de referência, como, por exemplo, o número total de genes/RNA/proteínas codificados em um genoma. A análise de enriquecimento identifica os termos que estão, estatisticamente, super-representados na lista de biomoléculas em análise, quando comparados ao valor de DF. Sugere-se que tais termos enriquecidos descrevem algum processo ou comportamento biológico subjacente importante (Tipney e Hunter, 2010).

Veja o estudo de caso a seguir. Consideremos o transcriptoma de uma espécie de planta submetida à desidratação radicular. Foi observado, após identificação dos RNAs contidos no referido transcriptoma, que 10% de todos os transcritos produzidos em resposta ao estresse em questão são codificadores de enzimas quinases. Esse valor é comparado com o DF para quinases no genoma da referida espécie. Foi constatado que apenas cerca de 1% (o valor de DF), de todos os genes do genoma da referida espécie são codificadores de quinases. Após tais informações, os valores (1% observado e 10% DF) são comparados, usando-se métodos estatísticos comuns (por exemplo, χ2, teste exato de Fisher, probabilidade binomial ou distribuição hipergeométrica). Caso haja indicação de diferença estatística nos conjuntos de dados analisados, é possível determinar se as quinases são enriquecidas na lista de transcritos. No presente exemplo, diremos que as quinases estão enriquecidas e, portanto, têm importantes funções na resposta da planta analisada sob estresse de desidratação radicular. O princípio da análise de enriquecimento se fundamenta no fato de que, caso haja preponderância quantitativa de um grupo de transcritos sobre outros também presentes nas amostras analisadas, tal grupo preponderante, provavelmente, está desempenhando papel significativo no processo sob análise.

Abordagens que utilizam a estratégia de análise de enriquecimento são um campo de pesquisa em ativo desenvolvimento. Em 2005, havia relatos de 14 ferramentas disponíveis; a pesquisa mais recente realizada nesse contexto de busca relata um crescimento significativo, apresentando com 68 ferramentas (Tipney e Hunter, 2010). Huang *et al*. (2009) as classificam como pertencentes a pelo menos uma das três categorias algorítmicas:

* Análise de enriquecimento singular (SEA, do inglês “*Singular Enrichment Analysis*”),
* Análise de enriquecimento de genes (GSEA, do inglês “*Gene Set Enrichment Analysis*”)
* Análise modular de enriquecimento (MEA, do inglês “*Modular Enrichment Analysis*”).

Todas as categorias apresentam seus prós e contras, não existindo um padrão ouro entre as ferramentas disponíveis. Para uma revisão abrangente do tema, consultar (Tipney e Hunter, 2010). É válido destacar, entretanto, que tais categorias apresentam o mesmo objetivo, sendo o *output* resultante apresentado de forma semelhante [seja por meio de gráficos de colunas simples (**Figura 8A**), gráficos hierárquicos (**Figura 8B**) ou heatmaps].



**Figura 8.** Apresentação de *output* de análise de enriquecimento, mostrando os processos enriquecidos para uma dada situação estudada. **A.** Via gráfico de barras, comparando os valores de uma lista personalizada (“*input list*”; barras de cor turquesa) com os valores de fundo (“*background/reference*”; barras de cor verde claro); **B.** Via categorização hierárquica (quanto mais vermelho é o retângulo, mais enriquecido é o termo associado).

Apesar da estatística de análises de enriquecimento está disponível e ser aplicada para diversas situações em genômica, transcriptômica e proteômica, não há relatos de sua aplicação no contexto dos ECRs em promotores. Tal desenvolvimento e execução seria de grande valia, produzindo informação de grande impacto científico, com derivações que atingiriam os seguintes pontos estratégicos:

* agregaria valor à fisiologia molecular dos organismos, elencando certas famílias de FTs mais associados (ou enriquecidas) nos promotores de famílias gênicas, vias metabólicas, grupos de transcritos específicos, dentre outros. Isso lançaria informação sobre a rede regulatória desses alvos;
* agregaria valor ao *status* informacional de genes com função desconhecida, apresentando possíveis FTs participantes de seu processo de regulação;
* auxiliaria na mineração de alvos com potencial biotecnológico, associando FTs, cuja participação em estresses é bem sedimentada, a genes cuja participação na resposta a condições não favoráveis ainda é desconhecida;

# **Objetivo geral**

Estabelecer as bases da plataforma PhytoProm, por meio da criação de um algoritmo em linguagem Python que auxilie nas matrizes....xxxxxxxx

# **Objetivos específicos**

# **Material e métodos**

# **Resultados**

# **10.** **Referências Bibliográficas**

Brazma, A., I. Jonassen, J. Vilo, and E. Ukkonen. 1998. “Predicting Gene Regulatory Elements in Silico on a Genomic Scale.” *Genome research* 8(11):1202–15.

Bussemaker, H. J., H. Li, and E. D. Siggia. 2001. “Regulatory Element Detection Using Correlation with Expression.” *Nature genetics* 27(2):167–71.

Chow, Chi-Nga *et al*. 2016. “PlantPAN 2.0: An Update of Plant Promoter Analysis Navigator for Reconstructing Transcriptional Regulatory Networks in Plants.” *Nucleic acids research* 44(D1):D1154-60.

Hehl, Reinhard and Lorenz Bulow. 2014. “AthaMap Web Tools for the Analysis of Transcriptional and Posttranscriptional Regulation of Gene Expression in Arabidopsis Thaliana.” *Methods in molecular biology (Clifton, N.J.)* 1158:139–56.

Higo, Kenichi, Yoshihiro Ugawa, Masao Iwamoto, and Tomoko Korenaga. 1999. “Plant Cis-Acting Regulatory DNA Elements (PLACE) Database: 1999.” *Nucleic Acids Research* 27(1):297–300. Retrieved (http://dx.doi.org/10.1093/nar/27.1.297).

Hoang, Xuan Lan Thi, Du Ngoc Hai Nhi, Nguyen Binh Anh Thu, Nguyen Phuong Thao, and Lam-Son Phan Tran. 2017. “Transcription Factors and Their Roles in Signal Transduction in Plants under Abiotic Stresses.” *Current genomics* 18(6):483–97.

Hsieh, Ming-Hsien *et al*. 2013. “Virus-Induced Gene Silencing Unravels Multiple Transcription Factors Involved in  Floral Growth and Development in Phalaenopsis Orchids.” *Journal of experimental botany* 64(12):3869–84.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. “Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources.” *Nature protocols* 4(1):44–57.

Hughes, J. D., P. W. Estep, S. Tavazoie, and G. M. Church. 2000. “Computational Identification of Cis-Regulatory Elements Associated with Groups of Functionally Related Genes in Saccharomyces Cerevisiae.” *Journal of molecular biology* 296(5):1205–14.

Jensen, L. J. and S. Knudsen. 2000. “Automatic Discovery of Regulatory Patterns in Promoter Regions Based on Whole Cell Expression Data and Functional Annotation.” *Bioinformatics (Oxford, England)* 16(4):326–33.

Joshi, Rohit *et al*. 2016. “Transcription Factors and Plants Response to Drought Stress: Current Understanding and Future Directions.” *Frontiers in plant science* 7:1029.

Khan, Aziz *et al*. 2018. “JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework.” *Nucleic acids research* 46(D1):D1284.

Kuska, B. 1998. “Beer, Bethesda, and Biology: How ‘genomics’ came into Being.” *Journal of the National Cancer Institute* 90(2):93.

Li, Yifeng, Chih-yu Chen, Alice M. Kaye, and Wyeth W. Wasserman. 2015. “The Identification of Cis-Regulatory Elements: A Review from a Machine Learning Perspective.” *Biosystems* 138:6–17. Retrieved (http://www.sciencedirect.com/science/article/pii/S0303264715001604).

Pilpel, Y., P. Sudarsanam, and G. M. Church. 2001. “Identifying Regulatory Networks by Combinatorial Analysis of Promoter Elements.” *Nature genetics* 29(2):153–59.

Rombauts, Stephane *et al*. 2003. “Computational Approaches to Identify Promoters and Cis-Regulatory Elements in Plant Genomes.” *Plant Physiology* 132(3):1162 LP-1176. Retrieved (http://www.plantphysiol.org/content/132/3/1162.abstract).

Roth, F. P., J. D. Hughes, P. W. Estep, and G. M. Church. 1998. “Finding DNA Regulatory Motifs within Unaligned Noncoding Sequences Clustered by Whole-Genome mRNA Quantitation.” *Nature biotechnology* 16(10):939–45.

Seo, Eunyoung, Doil Choi, and Choi. 2015. “Functional Studies of Transcription Factors Involved in Plant Defenses in the Genomics Era.” *Briefings in functional genomics* 14(4):260–67.

Song, Li *et al*. 2016. “Genome-Wide Transcriptome Analysis of Soybean Primary Root under Varying Water-Deficit Conditions.” *BMC genomics* 17:57.

Stormo, Gary D. 2013. “Modeling the Specificity of Protein-DNA Interactions.” *Quantitative biology (Beijing, China)* 1(2):115–30.

Trindade, L. M., R. van Berloo, M. Fiers, and R. G. F. Visser. 2005. “PRECISE: Software for Prediction of Cis-Acting Regulatory Elements.” *Journal of Heredity* 96(5):618–22. Retrieved (http://dx.doi.org/10.1093/jhered/esi094).

van Helden, J., B. Andre, and J. Collado-Vides. 1998. “Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies.” *Journal of molecular biology* 281(5):827–42.

Vilo, J. and K. Kivinen. 2001. “Regulatory Sequence Analysis: Application to the Interpretation of Gene Expression.” *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology* 11(6):399–411.

Vilo, J., A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen. 2000. “Mining for Putative Regulatory Elements in the Yeast Genome Using Gene Expression Data.” *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 8:384–94.

Wei, Kaifa and Huiqin Chen. 2018. “Global Identification, Structural Analysis and Expression Characterization of Cytochrome P450 Monooxygenase Superfamily in Rice.” *BMC genomics* 19(1):35.

Wingender, Edgar, Torsten Schoeps, Martin Haubrock, and Jurgen Donitz. 2015. “TFClass: A Classification of Human Transcription Factors and Their Rodent Orthologs.” *Nucleic acids research* 43(Database issue):D97-102.

Xie, Zhi, Shaohui Hu, Jiang Qian, Seth Blackshaw, and Heng Zhu. 2011. “Systematic Characterization of Protein-DNA Interactions.” *Cellular and molecular life sciences : CMLS* 68(10):1657–68.

Yadav, Satya P. 2007. “The Wholeness in Suffix -Omics, -Omes, and the Word Om.” *Journal of biomolecular techniques : JBT* 18(5):277.

Yao, Wenjing, Shengji Wang, Boru Zhou, and Tingbo Jiang. 2016. “Transgenic Poplar Overexpressing the Endogenous Transcription Factor ERF76 Gene Improves Salinity Tolerance.” *Tree physiology* 36(7):896–908.

Yilmaz, Alper *et al*. 2011. “AGRIS: The Arabidopsis Gene Regulatory Information Server, an Update.” *Nucleic acids research* 39(Database issue):D1118-22.

Zhu, J. and M. Q. Zhang. 2000. “Cluster, Function and Promoter: Analysis of Yeast Expression Array.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 479–90.