SUMÁRIO

1 INTRODUÇÃO A BIOLOGIA MOLECULAR DO FUNCIONAMENTO GÊNICO

Biologia molecular estuda a natureza química dos genes. Como a informação genética é codificada, replicada e expressa. Isto inclui os processos celular da transcrição, tradução e regulação do gene (PIERCE, 2012). Neste capitulo será revisado os processos dentro da biologia molecular, com ênfase na regulação de um gene no nível transcricional.

1.1 Ácidos nucleicos

Existe uma grande diversidade de seres vivos, mas a codificação das instruções de todos os organismos vivos estão escritas na mesma linguagem, a "linguagem"dos ácidos nucleicos. Estas são moléculas que exercem um importante papel nos organismos vivos, porque nelas, estão contidos o material genético. Um grande número de informação esta armazenado no material genético, instruções de todas as peculiaridades e funções dos organismos. É a partir do material genético que as células recebem instruções de quais proteínas sintetizar e em que quantidade. Essa informação é decifrada através do código genético, cuja a tradução resulta na síntese de proteínas (ZAHA, 2000). Existem dois tipos de ácidos nucleicos: ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA), ambos os ácidos são compostos por nucleotídeos. Os nucleotídeos são formados a partir de três componentes químicos: um açúcar, uma base nitrogenada e um ácido fosfórico (Figura 1.1). Os nucleotídeos estão ligados entre eles formando uma sequência linear (Figura 1.2).

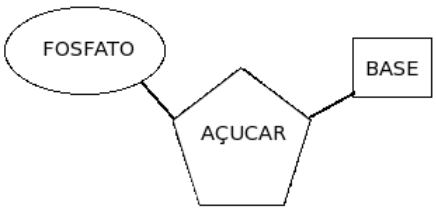

Figura 1.1: Componentes de um nucleotídeo

Existem duas importantes diferenças entre o RNA e o DNA, que são: o tipo de açúcar e

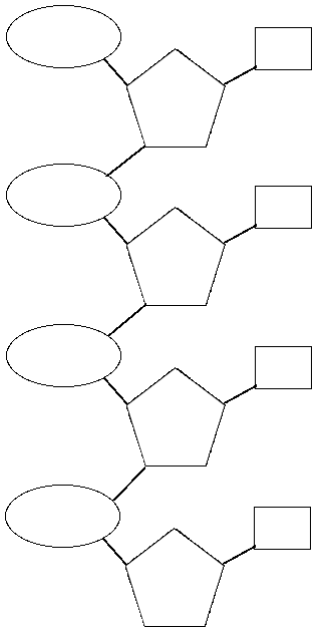

Figura 1.2: Sequência linear de nucleotídeos ligados

as bases nitrogenadas. O açúcar no DNA é a desoxirribose e no RNA a ribose (Figura 1.3). Em cada ácido nucleico são encontradas quatros bases nitrogenadas, sendo que três delas são compartilhadas entre o RNA e DNA são elas: adenina (A), guanina (G) e citosina (C). A base timina (T), é encontrada só no DNA, e a uracila (U), é encontrada só no RNA. Porém existe outra grande diferença entre o RNA e o DNA no nível estrutural. O RNA geralmente existe como uma única sequência de nucleotídeos, enquanto o DNA existe como duas sequências de nucleotídeos pareadas que formam um helicoide, conhecido como dupla hélice. Na Figura 1.4, podemos observar a estruturas dos dois ácidos. As bases nitrogenadas no DNA estão no interior da hélice, ligadas por pontes de hidrogênio formando pares de bases nitrogenadas (pb). Os únicos pares possíveis no DNA são: A ligado com T e C ligado com G (ZAHA, 2000).

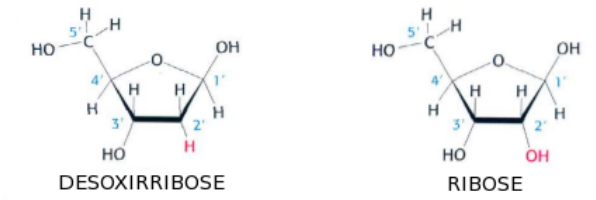

Figura 1.3: Tipos de açúcar encontrados nos ácidos nucleicos. (BERG; TYMOCZKO; STRYER, 2007, Adaptada))

Quanto a estrutura do RNA, ele pode ser dividido em várias classes conforme a sua funcionalidade na célula. A Tabela 1.1 mostra as diferentes classes de RNA e suas funcionalidades. O RNA apesar de ter apenas uma cadeia de nucleotídeos também tem bases complementares. Na

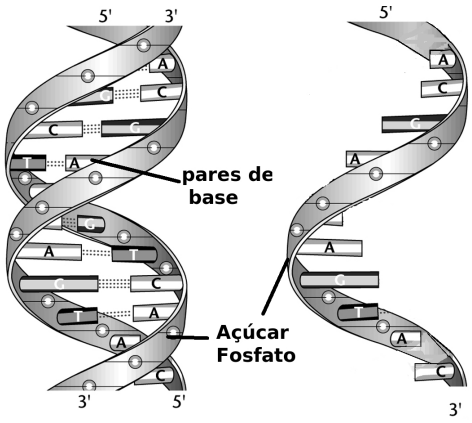

Figura 1.4: Estrutura do DNA e RNA. (HIGGS; ATTWOOD, 2005, Adaptada)

Classe de RNA	Função
RNA ribossômico (rRNA)	Componentes estrutural e funcional do ribossomos
RNA mensageiro (mRNA)	Carrega o código genético para a síntese de proteínas
RNA transportador (tRNA)	Transporta aminoácidos para o mRNA durante a síntese de proteínas
Small nuclear RNA (snRNA)	Processamento do pre-RNA
Small nucleolar RNA (snoRNA)	Processamento e montagem do rRNA
Small cytoplasmic RNA (scRNA)	Variável
MicroRNA (miRNA)	Inibe a tradução do mRNA
Small interfering RNA (siRNA)	Inicia a degradação de outras moléculas de RNA
Piwi-interacting RNA (piRNA)	Pouco sabe-se de sua função

Tabela 1.1: Funcionalidades das diferentes classes de moléculas de RNA

síntese de RNA, descrita com mais detalhes na Seção 1.3, as bases que compõem a sequência do RNA são o complemento das bases copiadas do filamento (sequência de nucleotídeos) do DNA modelo, com a substituição de T por U no RNA . As bases complementares no RNA são: A ligado com U e C ligado com G.

A Figura 1.5, mostra uma sequência de DNA e uma de RNA. Elas são comumente representas como uma palavra formada pelo alfabeto (A, G, C, T), para representações de DNA e (A, G, C, U), para representações do RNA. A leitura é feita da esquerda para a direita, no sentido indicado como 5'$\rightarrow$ 3'. Este tipo de representação torna fácil a visualização, assim como a manipulação a nível computacional, sendo amplamente utilizado em métodos *in silico* que envolvem o DNA e o RNA.

A interação entre o RNA e DNA ocorre quando é necessário a expressão de um gene. Uns

Figura 1.5: Reapresentação do RNA e DNA

dos primeiros passos para a expressão genética é um processo chamado transcrição (Seção 1.3). Neste processo ocorre a formação do RNA (síntese de RNA), a partir de um dos filamentos do DNA. A sequência de RNA formada é uma cópia exata de uma região do RNA. Uma parte desta região que é copiada pertence a um gene. Os genes são seguimentos de DNA podendo ter milhares de pares de bases. São eles que irão especificar o tipo de proteína a ser sintetizado. O processo da síntese de proteínas é conhecido como tradução (Seção 1.5). A Figura 1.6, apresenta cada passo deste conjunto de processos, que também é comumente conhecido como o dogma central.

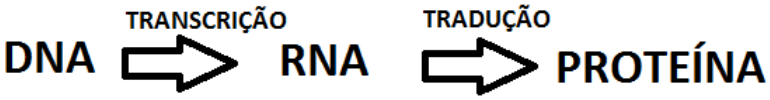

Figura 1.6: Principais passos da expressão de genética

1.2 Proteínas

Proteínas são os componentes principais de todos os seres vivos. Muitas proteínas são enzimas, que atuam como catalizadores biológicos em reações químicas na célula. Outras são componentes estruturais na célula, provendo suporte para membranas, filamentos, ossos, e cabelo. Algumas proteínas ajudam no transporte de substancias, e outras tem funções regulatórias, de comunicação ou de defesa (PIERCE, 2012).

As proteínas são compostas por uma sequência de aminoácidos. No total são encontrados pelo menos vinte aminoácidos nas proteínas.

Assim como os ácidos nucleicos a estrutura molecular das proteínas tem vários níveis de organização. A primeira estrutura da proteína é a sequência de aminoácidos. A segunda estrutura deriva da interação entre os aminoácidos que faz a estrutura dobrar-se, semelhante ao helicoide

do DNA. A terceira estrutura é a forma tridimensional formada pela ligação dos aminácidos na primeira estrutura. A quarta estrutura é a ligação de várias cadeias polipeptídicas[1]. A Figura 1.7 apresenta as quatro estruturas das proteínas.

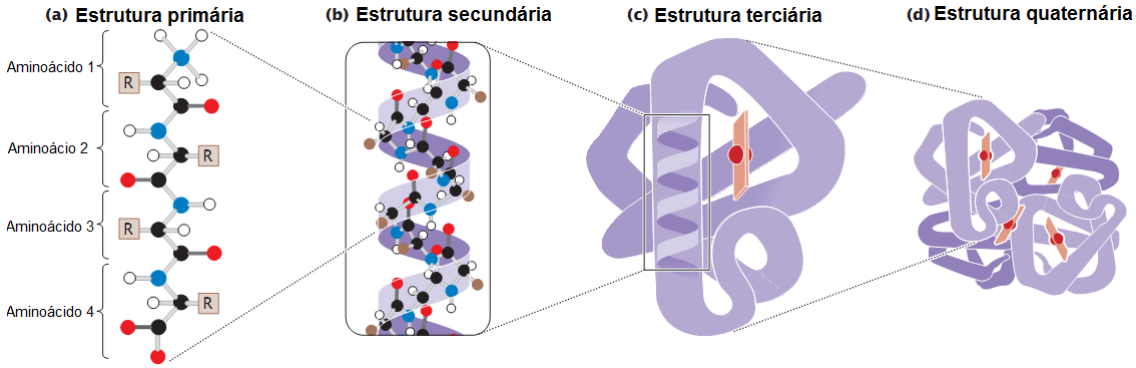

Figura 1.7: As quatro estruturas das proteínas: (a) A primeira estrutura de uma proteína é sua sequência de aminoácidos; (b) Interações entre aminoácidos faz com que a primeira estrutura se dobre; (c) A segunda estrutura se dobra ainda mais levando a terceira estrutura; (d) Dois ou mais cadeias polipeptídicas associam-se criando a quarta estrutura (PIERCE, 2012, Adaptada)

1.3 Transcrição

Transcrição é a síntese da molécula de RNA a partir de um molde de DNA. Neste processo uma pequena parte do DNA é copiado. Este tamanho pode variar entre poucos milhares de nucleotídeos, que em comparação com o DNA é muito pequeno, por exemplo uma molécula de DNA no cromossomo humano pode ser maior que 250 milhões de pares de nucleotídeos. Não são todos os genes em uma célula que são transcritos. Porque não é necessário que todas as instruções de funcionamento do organismo seja executada ao mesmo tempo ou na mesma célula (PIERCE, 2012). A transcrição é um processo altamente seletivo, uma vez que, os genes são transcritos somente quando seus produtos (proteínas) são necessários. Esta seletividade levanta um problema: como reconhecer os genes e transcrevê-los no tempo e local apropriado.

Para que ocorra a transcrição o primeiro passo é a abertura da hélice dupla do DNA, um dos filamentos do DNA servirá como modelo para a síntese de RNA. A sequência de nucleotídeos na cadeia de RNA é determinada pelo complemento do molde do filamento de DNA (Figura 1.8). Qual filamento servirá com molde isso depende do gene que será transcrito. O trecho do DNA que codifica a molécula de RNA e as sequências necessárias para sua transcrição, é chamado de unidade de transcrição. Um complexo de enzimas e proteínas, chamado de aparato

[1] Cadeias polipeptídicas são cadeias formadas por mais de vinte aminoácidos

de transcrição, reconhece uma unidade de transcrição. Aparece novamente o problema do paragrafo anterior, qual trecho de DNA será lido e como saber o início e a fim dessa leitura. Todas essas informações estão incluídas na sequência do DNA (PIERCE, 2012).

Dentro da unidade de transcrição há três regiões: um promotor, uma sequência de codificação do RNA, e um finalizador (Figura 1.9). O promotor é uma sequência que o aparato de transcrição reconhece e liga-se. Ele indica qual filamento de DNA será transcrito, em qual direção, e qual é o sítio de início da transcrição (TSS, do inglês *transcription start site*). O TSS é o local onde inicia-se a transcrição, ou seja o primeiro nucleotídeo a ser transcrito. A localização do promotor na unidade de transcrição é antes do TSS e ele não é transcrito em RNA. A segunda região na unidade de transcrição é a sequência de codificação do RNA, a sequência de nucleotídeos de DNA nessa região é transcrita para RNA. A terceira região é o finalizador, uma sequência de nucleotídeos que sinaliza onde é o fim da transcrição. Os finalizadores também fazem parte do RNA, a transcrição para após o finalizador ser copiado (PIERCE, 2012). Uma definição quanto a direção da transcrição e localização dos nucleotídeos aparece. A direção que o complexo de transcrição se move é chamada de *downstream*, a direção oposta é chamada de *upstream*. Quanto a localização dos nucleotídeos na sequência de DNA, ocorre da seguinte maneira. O primeiro nucleotídeo a ser transcrito (TSS) é numerado com +1, e todos os nucleotídeos localizados *downstream* do TSS são atribuídos números positivos. Nucleotídeos *upstream* do TSS são atribuídos números negativos. Não existe nucleotídeos numerados com 0.

Figura 1.8: RNA formado, complementar ao filamento modelo

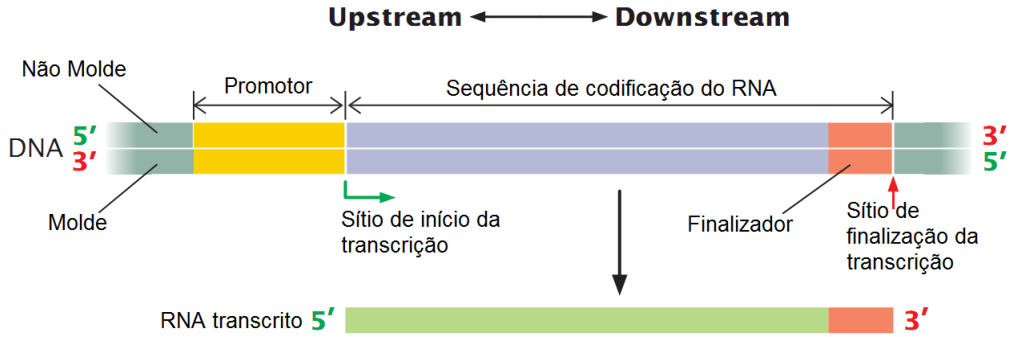

Figura 1.9: Unidade de transcrição, dividida em promotor, região codificante, e finalizador (PIERCE, 2012, Adaptada)

Para que ocorra a formação do RNA é necessário a ação de uma enzima que realiza a transcrição, ela é chamada de RNA polimerase. Ela juntamente com proteínas específicas formam o aparato de transcrição. Essas proteínas juntam-se e desconectam-se da RNA polimerase em diferentes níveis no processo de transcrição. Elas atuam como um reforço para a ação da RNA polimerase.

Nos organismos eucarióticos existem três RNA polimerase, chamadas de RNA polimerase I, RNA polimerase II, RNA polimerase III. As três enzimas são similares umas com as outras. A maior diferença entre elas é o tipo de RNA que elas transcrevem: RNA polimerase I transcreve rRNA; RNA polimerase II transcreve pre-mRNAs, snoRNAs, alguns miRNAs, e alguns snRNAs; e RNA polimerase III transcreve tRNA, pequenos rRNA,alguns miRNAs, e alguns snRNAs (Tabela 1.2). A RNA polimerase II transcreve a maioria dos genes, ela gera o RNA mensageiro, utilizado na formação das proteínas.

Tipo	Transcreve
RNA polimerase I	grandes rRNAs
RNA polimerase II	Pre-mRNA, alguns snRNAs, snoRNAs, alguns miRNAs
RNA polimerase III	tRNA, pequenos rRNA, snoRNAs, alguns miRNA

Tabela 1.2: RNA polimerase em eucarióticos (PIERCE, 2012, Adaptada)

O reconhecimento do promotor é feito por um conjunto de proteínas que ligam-se no promotor para que a RNA polimerase também possa se conectar no promotor. Este conjunto de

proteínas pode ser dividido em duas classes. A primeira composta pelos fatores de transcrição gerais, que juntamente com a RNA polimerase formam o aparato basal de transcrição. Este é formado por um grupo de proteínas que se ligam próximo do TSS, ele é suficiente para iniciar a transcrição em seu nível minimo. A outra classe é formada pelos fatores de transcrição específicos que se ligam a específicos segmentos de DNA, aumentando o nível de transcrição pela estimulação do aparato de transcrição na TSS (PIERCE, 2012). Neste trabalho será apresentado apenas o processo de transcrição com a RNA polimerase II, que transcreve os genes que codificam proteínas. Os promotores encontrados na RNA polimerase II, é dividido em duas partes: promotor principal e promotor regulatório.

O promotor principal está localizado *upstream* do gene e é o local onde o aparato basal de transcrição se conecta. O promotor principal inclui uma ou mais sequências consenso. Por exemplo um sequência consenso comum é a TATA-box, que é formada pela sequência TATAAA e é localizada aproximadamente -25 a -30 pares de base *upstream* do TSS. A Figura 1.10 mostra a TATA-box assim com sequências consensos adicionais.

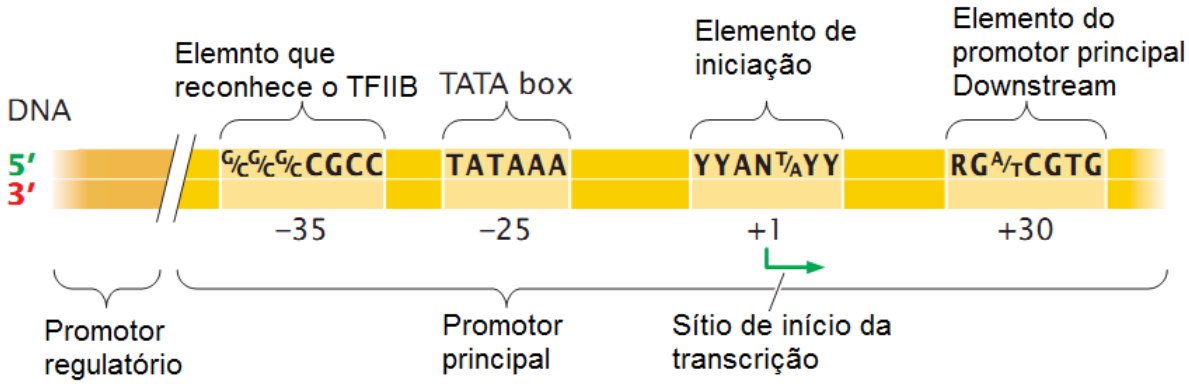

Figura 1.10: Promotor e algumas sequências promotoras, nem todas as sequências mostradas são encontradas em todos os promotores (PIERCE, 2012, Adaptada)

O promotor regulatório é localizado *upstream* do promotor principal. Neste promotor uma variedade de sequências consenso pode ser encontrada. Os fatores de transcrição específicos ligam-se a essas sequências e direto ou indiretamente faz contato com o aparato basal de transcrição, afetando a taxa que a transcrição é iniciada (Figura 1.11). Os fatores de transcrição específicos podem se conectar em sequências distantes chamadas de acentuadores. O DNA entre um acentuador e promotor forma uma dobra, e os fatores de transcrição ligados no acentuador podem interagir com o aparato de transcrição basal no promotor principal (SNUSTAD; SIMMONS, 2003).

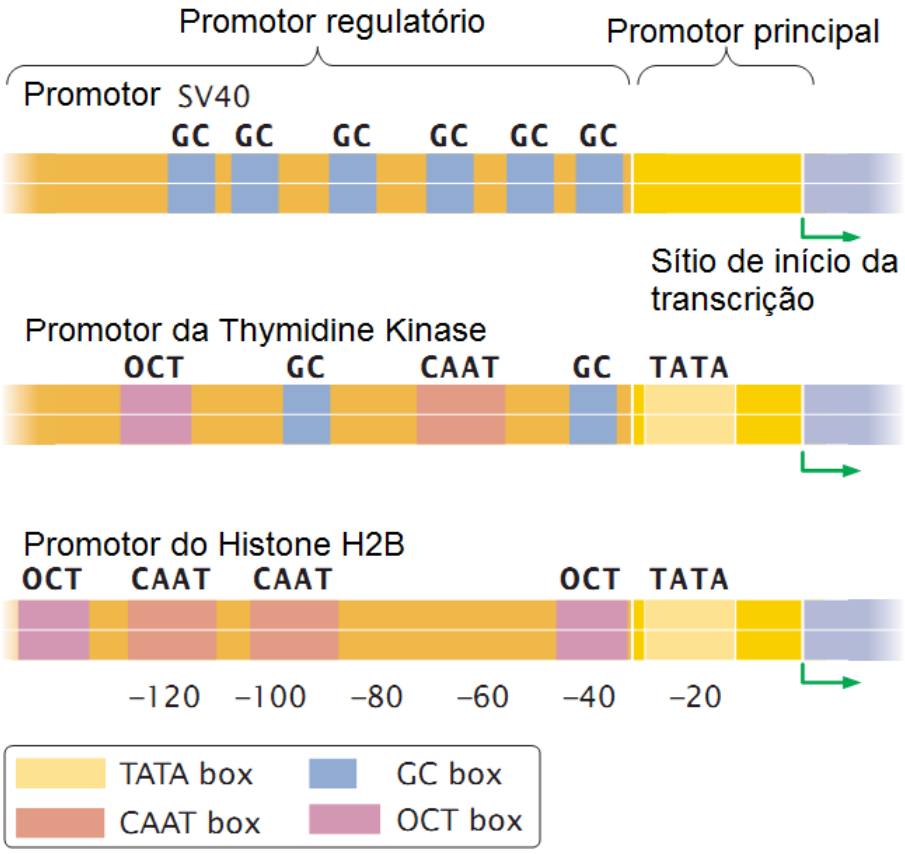

Figura 1.11: Três sequências promotoras de diferentes genes, note como as sequências consenso são ordenadas em diferentes combinações em cada gene(PIERCE, 2012, Adaptada)

Os fatores de transcrição gerais que formam o aparato basal de transcrição são: TFIIA, TFIIB, TFIID, TFIIE, TFIIF e TFIIH, onde TF é a sigla para *trancription factor* em inglês, a numeração II indica a RNA polimerase II e a letra identifica um fator de transcrição individual. O complexo formado é apresentado na Figura 1.12.

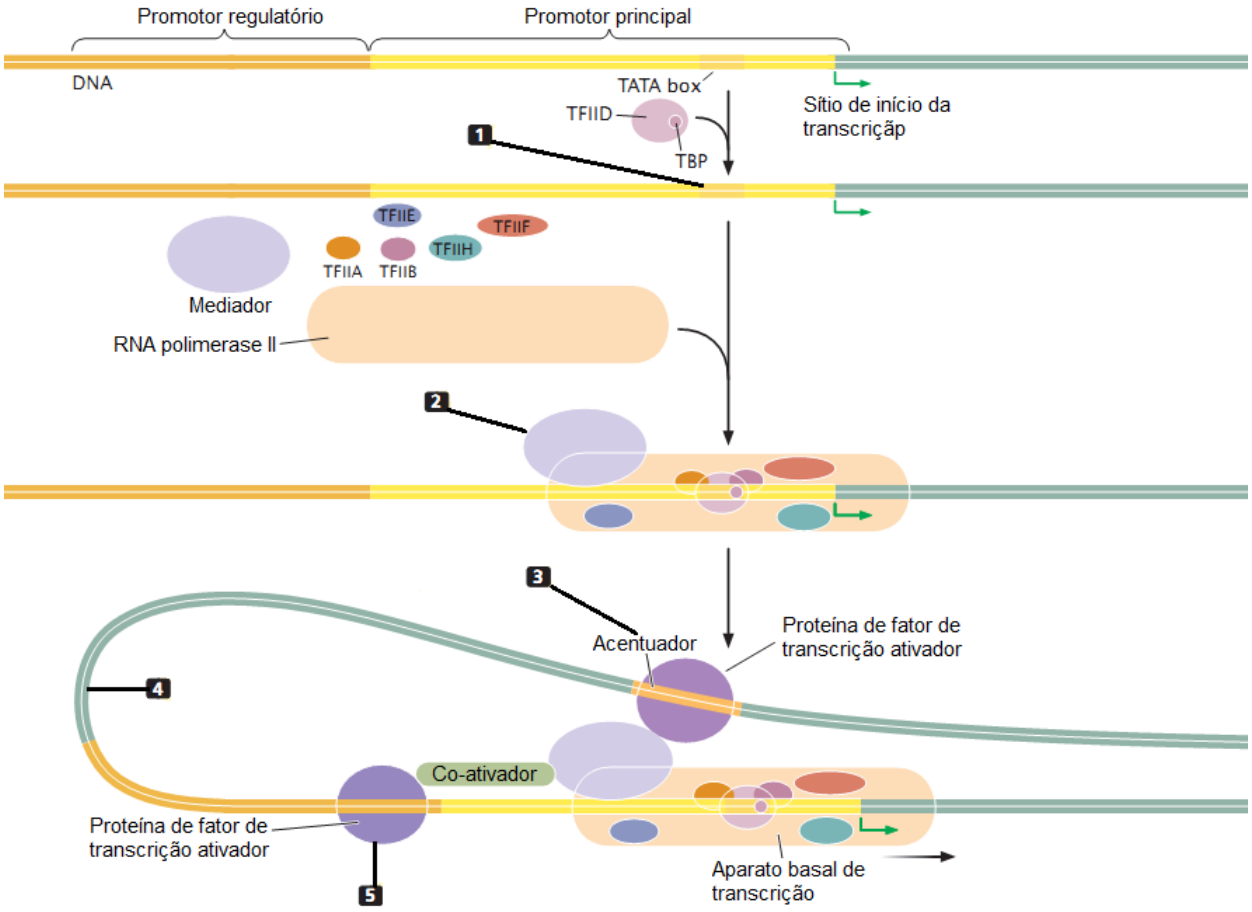

Figura 1.12: Início da transcrição com a RNA polimerase II: (1) TFIID conecta-se na TATA-box no promotor principal; (2) Fatores de transcrição e a RNA polimerase II conectam-se no promotor principal; (3) Proteínas de fator de transcrição específico conectam-se em sequências acentuadoras; (4) DNA faz uma volta, permitindo que proteínas se conectem nos acentuadores para interagir com o aparato basal de transcrição; (5) Proteínas de fatores de transcrição específico conectam-se nas sequências no promotor regulatório e interagem com o aparato basal de transcrição através de um mediador (PIERCE, 2012, Adaptada)

O primeiro passo para o inicio da transcrição é o TFIID se conectar na sequência de DNA TATA-box e separar parcialmente os filamentos. Outros fatores de transcrição se conectam em outras sequências consenso no promotor principal, na RNA polimerase e posicionam ela sobre o TSS.

Após a conexão da RNA polimerase e dos fatores de transcrição, formando o aparato de transcrição, a sequência de DNA é separada então a RNA polimerase move-se ao longo do

DNA na direção 5' para a 3', deixando o promotor e muitos fatores de transcrição, formando a cadeia de RNA que vai se alongando um nucleotídeo por vez, terminando com uma sequência de nucleotídeos exatamente complementar ao filamento de DNA usado como modelo (Figura 1.13). Após terminada a cópia a cadeia de RNA, juntamente com a RNA polimerase, se desconectam. A Figura 1.14 mostra o caminho percorrido da RNA polimerase durante a transcrição.

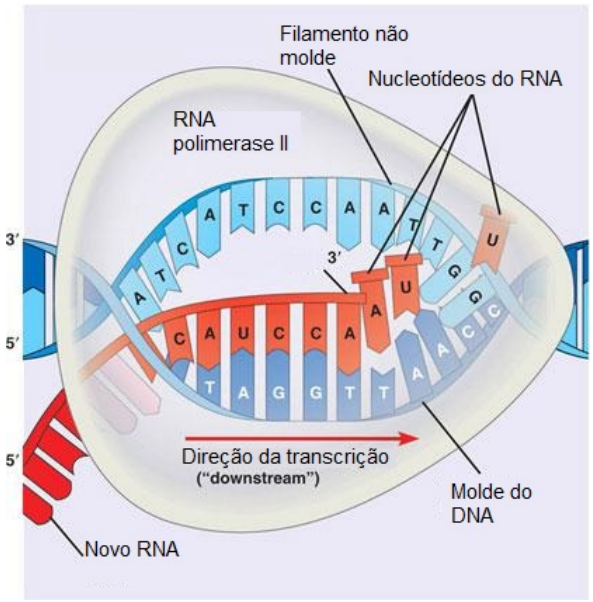

Figura 1.13: Formação do RNA através da RNA polimerase Fonte: http://www.bio.miami.edu/dana/250/250SS11_8.html

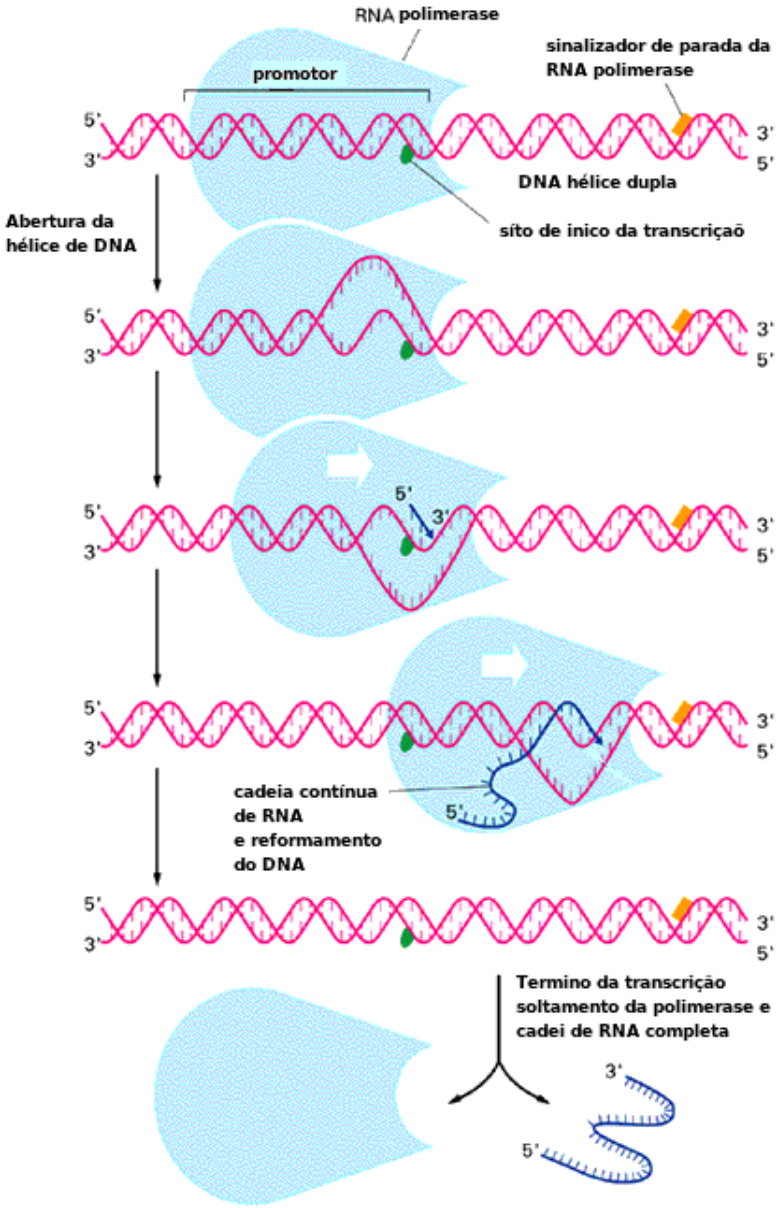

Figura 1.14: Movimento da RNA polimerase (HIGGS; ATTWOOD, 2005, Adaptada)

1.4 Código genético

Como já comentado na Seção 1.1, a informação codificada no DNA na forma de pares de nucleotídeos é a base para toda a diversidade, funcionalidade, e sobrevivência, nos seres vivos. Para utilizar a informação contida no DNA é necessário decodifica-la. O produto desta decodificação é o código genético.

O código genético é formado a partir da leitura do mRNA, ele é composto por palavras

de três nucleotídeos consecutivos. No total são permitidos $4^3 = 64$ possíveis códons [2]. Sendo que três destes códons são códons finalizadores, especificando o fim da tradução(Seção 1.5). Os outros 61 códons codificam aminoácidos. Existem vinte aminoácidos mas temos 61 códons para representa-los, isto significa que os aminoácidos são especificado por mais de um códon, com exceção do triptofano e da metionina, que são formados por apenas um códon (GRIFFITHS, 2000). Os códons que especificam o mesmo aminoácidos são chamados de sinônimos. A Figura 1.15 apresenta os aminoácidos e todos os possíveis códons. Já na Figura 1.16 mostra uma sequência de mRNA decodificada em códons.

PRIMEIRA LETRA	SEGUNDA LETRA: U	C	A	G	TERCEIRA LETRA
U	UUU phe	UCU ser	UAU tyr	UGU cys	U
	UUC phe	UCC ser	UAC tyr	UGC cys	C
	UUA leu	UCA ser	UAA parada	UGA parada	A
	UUG leu	UCG ser	UAG parada	UGG trp	G
C	CUU leu	CCU pro	CAU his	CGU arg	U
	CUC leu	CCC pro	CAC his	CGC arg	C
	CUA leu	CCA pro	CAA gln	CGA arg	A
	CUG leu	CCG pro	CAG gln	CGG arg	G
A	AUU ile	ACU thr	AAU asn	AGU ser	U
	AUC ile	ACC thr	AAC asn	AGC ser	C
	AUA ile	ACA thr	AAA lys	AGA arg	A
	AUG met	ACG thr	AAG lys	AGG arg	G
G	GUU val	GCU ala	GAU asp	GGU glu	U
	GUC val	GCC ala	GAC asp	GGC glu	C
	GUA val	GCA ala	GAA glu	GGA glu	A
	GUG val	GCG ala	GAG glu	GGG glu	G

Figura 1.15: O código genético formado por 64 códons Fonte: http://www.biomania.com.br/bio/conteudo.asp?cod=1238

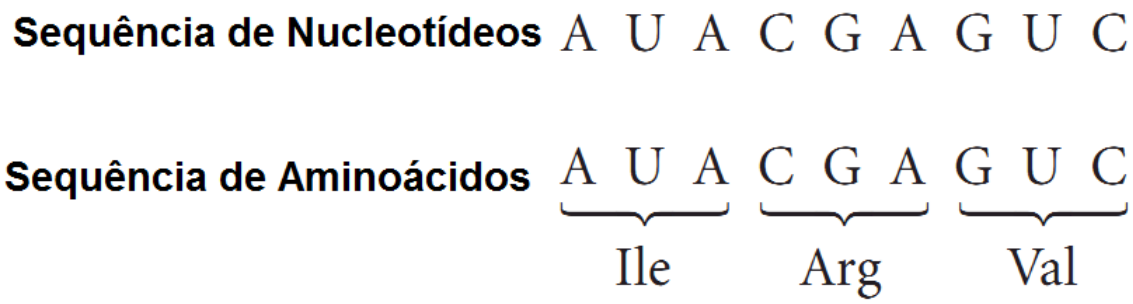

Figura 1.16: Decodificação do mRNA para aminoácidos

[2] Códons é a trinca formada pelos nucleotídeos consecutivos

1.5 Tradução

A tradução é processo da síntese da proteína a partir do mRNA. Diferente da transcrição a tradução não ocorre núcleo, mas sim no citoplasma da célula onde estão localizados o ribossomos. Os ribossomos são formados de duas subunidades, uma grande e uma pequena. Na tradução mRNA é decodificado para produzir uma sequência de polipeptídeo de acordo com a trinca no código genético. Ou seja utilizando uma cadeia de mRNA será formado a síntese de uma cadeia de aminoácido que por sua vez formará uma proteína (SNUSTAD; SIMMONS, 2003). O processo da tradução pode ser dividido em quatro partes:

1. **Ativação**

   Um aminoácido é ligado ao complemento de sua trinca no tRNA. Quando há uma ligação entre um aminoácido e um tRNA, este é chamado de "carregado".

2. **Iniciação**

   Nesta etapa os componentes necessários para a tradução são montados no ribossomo, com a ajuda dos fatores de iniciação, que são proteínas que auxiliam o processo.

3. **Alongamento**

   Nesta fase aminoácidos são juntados um por vez. Os tRNA carregados se conetam formando uma sequência de aminoácidos (cadeia polipeptídica).

4. **Término**

   O último processo, na qual a síntese de proteína para em um códon finalizador e os componentes da tradução são soltos do ribossomo.

Durante ou depois deste processo a cadeia de polipeptídeo assumi as estruturas secundária, terciária e quaternária da proteína. A figura 1.17 resume o processo da tradução.

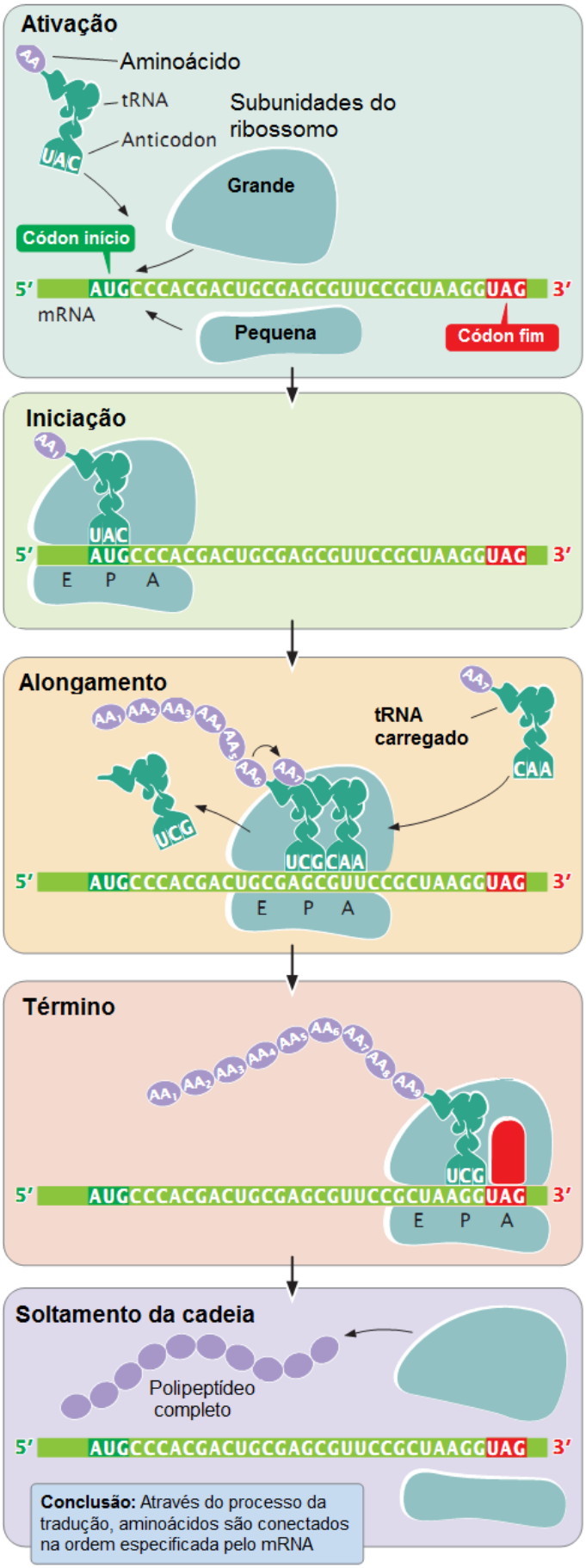

Figura 1.17: As quatro fases da tradução

1.6 Regulação no início da transcrição

Existem vários níveis de controle da expressão de um gene. Entre estes processão então : a transcrição, alteração da cromatina, alteração da estrutura do DNA, processamento e degradação do RNA, processos que afetam a tradução ou a modificação de proteínas. Nesta seção da continuidade a Seção 1.3, detalhando a ação dos fatores de transcrição específicos, a classificação que os fatores de transcrição recebem segundo suas funções, outras proteínas envolvidas que interagem com eles, e por último genes que têm sua regulação combinada.

Os fatores de transcrição específicos são divididos em duas classes: os ativadores, estimulam a transcrição; e os repressores inibem a transcrição. Os ativadores estimulam e estabilizam o aparato basal de transcrição. Eles podem atuar diretamente com o aparato basal de transcrição ou indiretamente através das proteínas co-ativadoras. Os ativadores podem se conectar em uma sequência de base, geralmente consenso, no promotor regulatório ou em um acentuador. No promotor regulatório há diferentes sequências consenso, em que diferentes fatores ativadores se conectam. Essa sequências consenso são chamadas de elementos regulatórios, e elas podem formar várias combinações (Figura 1.11), então cada promotor é regulado por uma combinação única de fatores de transcrição ativadores (PIERCE, 2012). Múltiplos elementos de regulatórios formam os CRMs (do inglês, *cis-elements modules*), que integra a conexão de vários fatores de transcrição resultando em um controle combinatório, e em um padrão específico da expressão de um gene (PRIEST; FILICHKIN; MOCKLER, 2009). As funções dos elementos regulatórios e dos CRMs são essenciais as respostas celulares a estímulos (PRIEST; FILICHKIN; MOCKLER, 2009).

A outra classe de fatores de transcrição são os repressores, eles se conectam em elementos regulatórios no promotor regulatório ou em sequências distantes chamadas de silenciadores, que semelhantes aos acentuadores eles são independente de posição e orientação (PIERCE, 2012). Os repressores competem com os ativadores em três situações:

1. Se uma região é ocupada por um ativador, a transcrição é ativada, mas se um repressor ocupa o espaço, não há ativação.

2. Um repressor conecta-se próximo a um ativador e impede o ativador interagir com o aparato de transcrição.

3. Um repressor interfere diretamente na montagem do aparato de transcrição, bloqueando a iniciação da transcrição.

1.6.1 Regulação combinada

Alguns genes são ativados pelos mesmos estímulos, eles têm em comum os mesmos elementos regulatórios no promotor ou no acentuador. Estes elementos regulatórios são chamados de elementos de respostas, são sequências pequenas que geralmente têm uma sequência consenso (Tabela 1.3). Os elementos de resposta são locais de ligação para os fatores de transcrição, que quando ligados aumenta o nível da transcrição. O mesmo elemento de resposta pode estar presente um vários genes permitindo que múltiplos genes sejam ativados pelo mesmo estímulo. Um estímulo pode ser um estresses como mudança na temperatura, mudança hormonal entre outros. Na próxima seção detalharemos um importante fator de transcrição (*dehydration responsive element binding proteins*) que se liga no elemento de resposta *dehydration responsive element*.

Elemento de resposta	Tipo de resposta	Sequência consenso
Heat-shock element	Calor e outros estresses	CNNGAANNTCCNNG [3]
Glucocorticoid response element	Glococorticoids	TGGTACAAATGTTCT
Phorbol ester response element	Phorbal esters	TGACTCA
Serum response element	Serum	CCATATTAGG
Dehydration responsive element	mudanças de temperatura, alta salinidade, e seca	A/GCCGAC [4]

Tabela 1.3: Alguns elementos de resposta

1.7 Dehydration responsive element binding proteins (DREB)

No amplo conjunto de fatores de transcrição, existem aqueles que quando ligados nos elementos de resposta irão ativar as respostas da célula a estresses abióticos. O estresse abiótico afeta diversos organismos, mas em especial os organismos vegetais que são dependentes de fatores ambientais, são os mais afetados. Nesta seção apresentaremos o *dehydration responsive element binding proteins* (DREB) um importante fator de transcrição nas plantas.

O DREB ativa genes que estão relacionados com a resposta da célula a estímulos abióticos, com a ativação destes genes, a planta se adapta as condições adversas a sua sobrevivência, através de reações bioquímicas e físicas que ocorrem na planta. Os estresses abióticos que mais

afetam as plantas são: seca, alta salinidade e mudanças de temperatura. O estresse abiótico atrapalha a sobrevivência e consequentemente a produção de grãos como a soja, arroz, milho e o trigo.

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

O elemento de resposta que se conecta no DREB é chamado de *dehydration responsive element* DRE, e ele é formado pela sequência A/GCCGAC (NAKASHIMA; ITO; YAMAGUCHI-SHINOZAKI, 2009). Entretanto para que o DRE seja funcional ele tem que estar acompanhado de elementos regulatórios, mas a especificidade desses elementos regulatórios é baixa, fazendo que muitos desses variem de gene para gene (ZHANG et al., 2005). Portanto, em uma busca computacional de genes alvos do DREB (ou outro fator de transcrição), não basta procurar o DRE (ou outro elemento resposta), na região promotora, uma vez que outros elementos regulatórios também participam da regulação, isto torna a busca por genes alvo uma difícil tarefa, uma vez que há uma variação nos elementos regulatórios de um gene para outro.

Segundo (AGARWAL et al., 2006), o entendimento do DREB na regulação de um gene é de grande importância para o desenvolvimento de plantas tolerantes a estresses. Já que, estresses abióticos e bióticos influenciam negativamente na sobrevivência e na larga produção de grãos. Culturas como soja, arroz e trigo que são amplamente usadas na alimentação mundial são prejudicadas pelos estresses que muitas vezes impedem uma alta produtividade.

2 BUSCA DE GENES ALVOS DE FATORES DE TRANSCRIÇÃO

Para decifrar a regulação gênica de um organismo, e consequentemente ter um domínio na manipulação genética do organismo, muitas metodologias *in vitro* e *in silico* foram propostas. Das metodologias *in silico*, algumas focam na busca de genes alvos de fatores de transcrição, e outras na classificação da funcionalidade de um gene desconhecido a partir de um gene conhecido, busca de elementos regulatórios, construção da rede de regulação, entre outras. A seguir serão apresentados algumas metodologias para a busca de genes alvos de fatores de transcrição. Esta é uma tarefa importante, porque identificando o fator de transcrição associado ao gene é possível saber qual a funcionalidade do gene. A descoberta da funcionalidade de um gene e quais os fatores de transcrição estão associados com a transcrição deste, tem várias utilidades dependendo do organismo. Por exemplo em uma planta abre a possibilidade de fazer um melhoramentos genético. Em bactérias desenvolver drogas que inibem genes essenciais para a seu desenvolvimento. Existem inúmeras aplicações variando em cada organismo.

2.1 Estado da arte

Para inferir genes alvos do DREB na *Arabidopsis* (WANG et al., 2009), criaram uma estratégia computacional, que combina a análise de elementos regulatórios e aprendizagem de máquina.

Eles utilizaram como conjunto de dados: sequências da região promotora de genes alvos do DREB identificados experimentalmente, estes são identificados como genes mestres (MGs, do inglês *Master genes*); *DRE frame sequences* (DFSs), fragmentos de DNA com 206 pb (pares de bases), retirados das regiões promotora de MGs, contendo a subsequência consenso (A/GCCGAC) que é a região onde o DREB se conecta em um gene, identificada pelos autores como DRE-motif [1]; *Non-DRE frame sequences* (nDFS) fragmentos de DNA com 206 pb,

[1] *motif* são regiões conservadas no DNA, assim todos os elementos regulatórios no DNA são *motif*, e não todos os *motifs* são elementos regulatórios

retirados das regiões promotora de gene aleatórios, com um DRE-motif inserido artificialmente na região central. No total foram encontrados 48 DFSs de regiões promotora de MGs, que foram considerados como dados positivos, e 1000 nDFSs como dados negativos.

Após a seleção do conjunto de dados, foi construído um classificador SVM para categorizar DFSs e nDFSs. O vetor de características utilizado no SVM, foi um vetor contendo hexâmeros (pequenos fragmentos de DNA com 6 nucleotídeos), que foram selecionados através do algoritmo HexDiff (CHAN; KIBLER, 2005). Este algoritmo computa o valor da frequência de um hexâmero no conjunto positivo $F_p(h)$ e negativo $F_n(h)$ em ambos filamentos de DNA. Depois de calculada a frequência para cada hexâmero em ambos os conjuntos de dados foi calculado o valor da razão $R(h)$, como:

$$R(h) = \frac{F_p(h)}{F_n(h)} \tag{2.1}$$

Os hexâmeros com a razão maior que um limite estabelecido, foram colocados em um vetor H_d, que é o vetor de características da SVM. Para cada hexâmero em H_d que pertence ao conjunto de DFS foi atribuído o rótulo (+1), e os de nDFS foi atribuído (-1). A função Kernel utilizada no classificador SVM foi a *Radial Basis Function* (RBF). Para o treinamento do classificador, visto que havia uma disparidade grande entre os dados positivos e negativos, foram usadas apenas 100 amostras negativas e todas as 48 amostras positivas. Já na classificação de DFSs e nDFSs foram utilizadas as 1000 amostras negativas, os dados positivos passaram por um processo de reamostragem, uma vez que eles estavam em um quantidade bem menor (48 DFSs) em relação aos dados negativos, assim o conjunto de dados positivos aumentou para 500 amostras.

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

Segundo os autores, apenas encontrar regiões nos promotores de um gene contendo o consenso DREB, não é suficiente para inferir este gene como alvo do DREB, devido as perdas de características do DRE-motif. Vários estudos, apontam que elementos regulatórios distribuídos na região promotora influenciam na ligação de um fatore de transcrição e seu gene alvo. O que pode ser entendido que, na PR de genes alvos do DREB não somente o DRE-motif mas também outros elementos regulatórios, influenciam na conexão de um DREB na sequência. Estes outros

elementos regulatórios agem como auxiliares para promover a conexão do DREB nos genes alvos. Com a utilização do HexDiff é selecionado regiões conservadas, que podem ser partes de elementos regulatórios que influenciam na ação do DREB, portanto um DFS será formado além da região consenso, também por sub-regiões conservadas.

(HOLLOWAY; KON; DELISI, 2008) projetaram uma metodologia utilizando aprendizagem de máquina para a predição de genes alvos de fatores de transcrição específicos e análise da rede regulatória do *Saccharomyces cerevisiae*, uma levedura muito utilizada em estudos genéticos por ter um genoma pequeno e simples comparado a outros organismos.

Para a classificação dos genes alvos de fatores de transcrição, eles utilizaram uma SVM, porque segundo os autores, os conjuntos de dados genômicos têm uma alta dimensionalidade, podendo chegar a milhares ou dezena de milhares de características numéricas para descrever um gene. Assim muitos algoritmos classificadores, podem ter um baixo desempenho com um número alto de características, ao contrário da SVM que tem um bom desempenho com dados de alta dimensionalidade.

Como dados de entrada positivo para o treinamento da SVM, foram pegos genes com elementos regulatórios conhecidos, que se ligam a determinados fatores de transcrição que também são conhecidos. Os elementos regulatórios, seus respectivos genes e os fatores de transcrição, foram selecionados por meio de publicações na literatura. O conjunto de entradas negativo, foi escolhido de subconjuntos de genes que têm um alto p-valor, consequentemente com uma probabilidade baixa de ter uma ligação nos fatores de transcrição utilizados.

Depois de selecionado o conjunto negativo são construídos 50 classificadores para cada fator de transcrição, utilizando diferentes subamostras do conjunto negativo, com o mesmo tamanho da amostra positiva. É utilizado 50 classificadores para cada fator de transcrição, porque há uma pequena probabilidade de um gene ser associado incorretamente ao conjunto negativo. Com os 50 classificadores este inoportuno é suavizado.

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

As características do classificador são selecionadas aplicando o algoritmo SVM *recursive feture elimination* (SVM-RFE), que otimiza o vetor $\mathbf{w}$ da SVM, para conter componentes altas, que são melhores para separar as classes positivas e negativas de dados. O processo de seleção de características, usando o SVM-RFE, é repetido até atingir o numero desejado de características, que é 1500, este numero é escolhido porque, quando são usadas 1500 características a

medida de precisão é aproximadamente 85%, que é uma precisão consideravelmente boa. Estas características são escolhidas para cada fator de transcrição e são guardadas durante a avaliação dos 50 classificadores. No final obtiveram um grande conjunto de características para um fator de transcrição, lembrando que os elementos desse conjunto são subconjuntos de características. Então foram escolhidos os 40 características com maior valor, de cada subconjunto de características, que foram guardas em uma lista e é contada quantas vezes cada característica apareceu. Esta lista foi rearranjada, posicionando os elementos que tem uma frequência de aparição maior no topo, resultando o conjunto final de características. Essas características inclui um conjunto diverso de dados incluindo sequências promotoras, medidas de expressão de um gene, conservação filogenética de elementos de sequências, sobre-representação de sequências promotoras, temperatura de fusão de promotores, e outras.

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

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

(LAN et al., 2007) desenvolveu um método de classificação de genes segundo suas funcionalidades, mais especificamente de genes que estão envolvidos na resposta das plantas a estresses.

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

condições experimentais. O valor retornado por um classificador tem que ser maior que o limite estabelecido.

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

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

Os classificadores foram treinados com todas as dimensões geradas pelo PCA e pela dimensão original com exceção do KNN, no caso do KNN é usado diferentes valores de K na dimensão original, onde K são os K vizinho mais próximos de um gene. Então foi escolhida a dimensão em que cada algoritmo obteve melhores resultados (no caso do KNN o melhor K). Os classificadores com a sua melhor dimensão, foram combinados em um só classificador, onde o valor de discriminação do classificador combinado é uma combinação linear dos discriminastes dos classificadores individuais. Como esperado o classificador combinado obteve melhores resultados que os classificadores individuais.

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

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

O HexDiff é um tipo de algoritmo de aprendizado de máquina, e foi projetado para discriminar dois tipos de sequências de DNA: CRM, e non-CRM(não agrupamento de elementos

regulatórios). Para fazer a classificação é necessário um conjunto de dados de treinamentos, que é obtido através de conhecidos CRMs, que são colocados no conjunto positivo de treinamento, os não conhecidos, os non-CRMs, são inseridos no conjunto negativo. Os dados para teste foram pegos de conhecidos CRMs da *Drosophila*. Foram encontrados 16 genes que, contém um total de 52 CRMs. Após a seleção dos dados foi calculada a frequência de cada hexâmero, no conjunto negativo $f_p(h)$ e positivo $f_n(h)$, para calcular a razão $R(h)$:

$$R(h) = \frac{F_p(h)}{F_n(h)} \tag{2.2}$$

Os hexâmeros mais comuns em CRMs do que em non-CRMs, obtiveram um alto valor de $R(h)$ e foram armazenados no conjunto H_d.

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

Para a avaliação foi usado a abordagem *leave-one-out cross-validation* (LOOCV), onde dos 16 genes encontrados 15 são treinados e 1 é usado como teste, este processo é feito até que todos os 16 genes sejam usados no conjunto de teste. A precisão das previsões do modelo foi medida com a correlação de Matthew.

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

Outro trabalho foi o de (ZHANG et al., 2005), que projetaram uma aplicação para encontrar genes alvos de fatores de transcrição, na *Arabidopsis thaliana* ,que são induzidos por ácido abscísico(ABA, do inglês *abscisic acid*) e por estresses abióticos.

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

Para encontrar os genes alvos, primeiramente foi calculada a pontuação de cada subsequência de tamanho w em uma determinada sequência, baseado em padrões com relevâncias biológicas (*motifs*) e em um modelo de *background*. Um *motif* W de tamanho w é representado

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

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

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

O primeiro passo foi, identificar, classificar, e associar as regiões não codificantes dos genes, isto foi feito com as informações contidas em arquivos GFF (do inglês, *General Feature Format*) extraídos de bancos de dados públicos. Foram retiradas as regiões codificantes porque é improvável que há elementos regulatórios nestas regiões e por elas serem impróprias para buscas filogenéticas apresentando uma alta conservação. O segundo passo consistiu em construir uma PWM (do inglês, *Position Weight Matrix*) para um conjunto de elementos regulatórios definidos experimentalmente, neste trabalho foram usados os elementos regulatórios dos fatores de transcrição TTX-3 e CEH-10. Para a construção da PWM foi usado o software HMMER (EDDY, 1998), este software utiliza o modelo oculto de Markov para gerar a PWM. A PWM resultante é usada como entrada no CisOrtho, que faz uma busca nas sequências não codificantes, atribuindo uma pontuação para cada bloco de subsequência de tamanho n (chamado de

janela), onde o objetivo é encontrar subsequências com a maior pontuação *N*, e com no máximo *D* subsequências, onde *N* e *D* são definidos pelo usuário. No quarto passo foi feita a análise filogenética, foram utilizados conjuntos de mapeamentos ortólogos[2] entre *C. elegans* e *C. briggasae* que provê pares de subsequências ortólogas[3]. Os pares de subsequências que apresentam uma alta pontuação são guardados, como um par de subsequência. O último passo classifica os pares de subsequências de acordo com as suas pontuações e *mismatches*[4].

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

[2]Sequências conservadas de diferentes organismo, que tem o mesmo ancestral
[3]sequências derivadas de um mesmo ancestral
[4]Bases diferentes, em uma mesma posição, neste caso é comparado as bases do *C. elegans* e *C. briggasae*

2.2 Tabela comparativa

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

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

2.3 Discussão

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

Observamos o importante papel que algoritmos de aprendizado de máquina tem na busca de genes alvos, estes algoritmos são usados em pelo menos uma das etapas das metodologias

apresentadas. Em (HOLLOWAY; KON; DELISI, 2008) e (LAN et al., 2007), é discutido que os dados genômicos apresentam uma alta dimensionalidade, o que nos leva a concluir que a predominância do SVM nas metodologias é devido a facilidade que SVM tem em lidar com alta dimensionalidades e os bons resultados obtidos mediante estas condições.

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

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

REFERÊNCIAS BIBLIOGRÁFICAS

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

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

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

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

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

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

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

GRIFFITHS, A. J. F. *An introduction to genetic analysis*. W.H. Freeman, 2000. ISBN 071673771. Disponível em: <http://www.worldcat.org/isbn/071673771>.

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

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

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

NAKASHIMA, K.; ITO, Y.; YAMAGUCHI-SHINOZAKI, K. Transcriptional Regulatory Networks in Response to Abiotic Stresses in Arabidopsis and Grasses. *Plant Physiology*, v. 149, n. 1, p. 88–95, jan. 2009. ISSN 0032-0889. Disponível em: <http://dx.doi.org/10.1104-/pp.108.129791>.

PIERCE, B. A. *Genetics : a conceptual approach*. W.H. Freeman, 2012. ISBN 1429232528. Disponível em: <http://www.worldcat.org/isbn/1429232528>.

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

SNUSTAD; SIMMONS, M. J. *Principles of genetics*. John Wiley & Sons, 2003. ISBN 0471441805. Disponível em: <http://www.worldcat.org/isbn/0471441805>.

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

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

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