Estratégias computacionais para a busca de genes alvos de *dehydration responsive binding proteins* (DREBs) na soja

Josué Crispim

13 de maio de 2011

Sumário

1 Transcrição

2 Elementos regulatórios e Fatores de transcrição

3 Dehydration Responsive Binding Proteins(DREBs)

4 Busca de elementos regulatórios

5 Identificação de genes alvos de DREBs

6 Conclusão

Início da regulação de um gene

- Um dos primeiros passos para a expressão de um gene é a transcrição.
- Onde diversos fatores podem influenciar a indução ou a repressão da expressão.

Ácidos nucleicos

- Os ácidos nucleicos são importantes moléculas que contem o material genético da célula.
- Uma analogia de sistemas biológicos com sistemas de computadores é pensar nos ácidos nucleicos como um código objeto de um programa, onde este código é descifrado pelo o sistema operacional (a célula) que irá tomar as devidas ações. No caso de células a ação é a produção de proteínas.

Ácidos nucleicos

- A composição química dos ácidos nucleicos é: um açúcar, uma base nitrogenada e um ácido fosfórico.
- Eles são ligados formando uma sequência linear.

Ácidos nucleicos

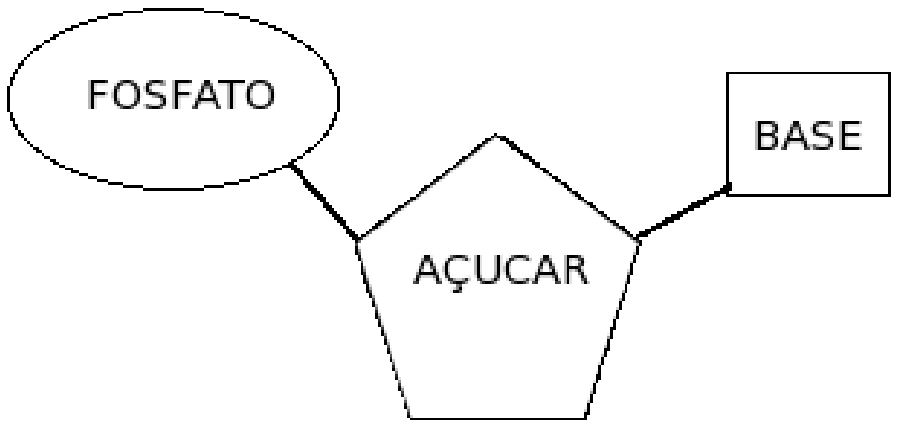
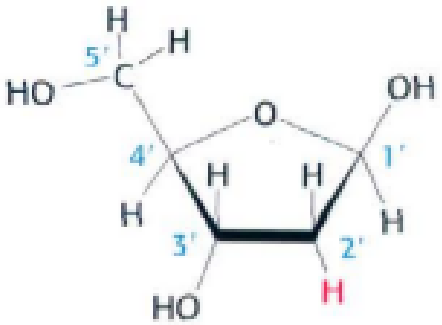
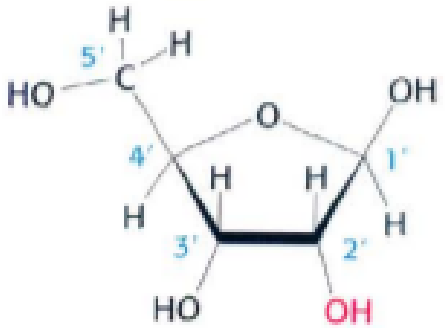


Figura: Componentes de um nucleotídeo

Ácidos nucleicos



DESOXIRRIBOSE



RIBOSE

Figura: Tipos de açúcar encontrados nos ácidos nucleicos. [**?**, Adaptada])

Ácidos nucleicos

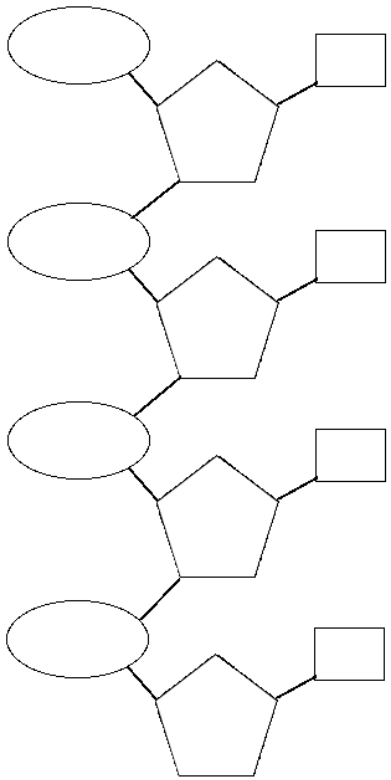


Figura: Sequência linear de nucleotídeos ligados

Ácidos nucleicos

Existem dois tipos de ácidos nucleicos:

1. Ácido desoxirribonucleico (DNA)
   - açúcar: desoxirribose
   - bases: A,**T**, G e C
   - estrutura: duas sequências complementares pareadas formando um helicoide
2. Ácido ribonucleico (RNA)
   - açúcar: ribose
   - bases: A, **U**, G e C
   - estrutura: única sequência de nucleotídeos

Ácidos nucleicos

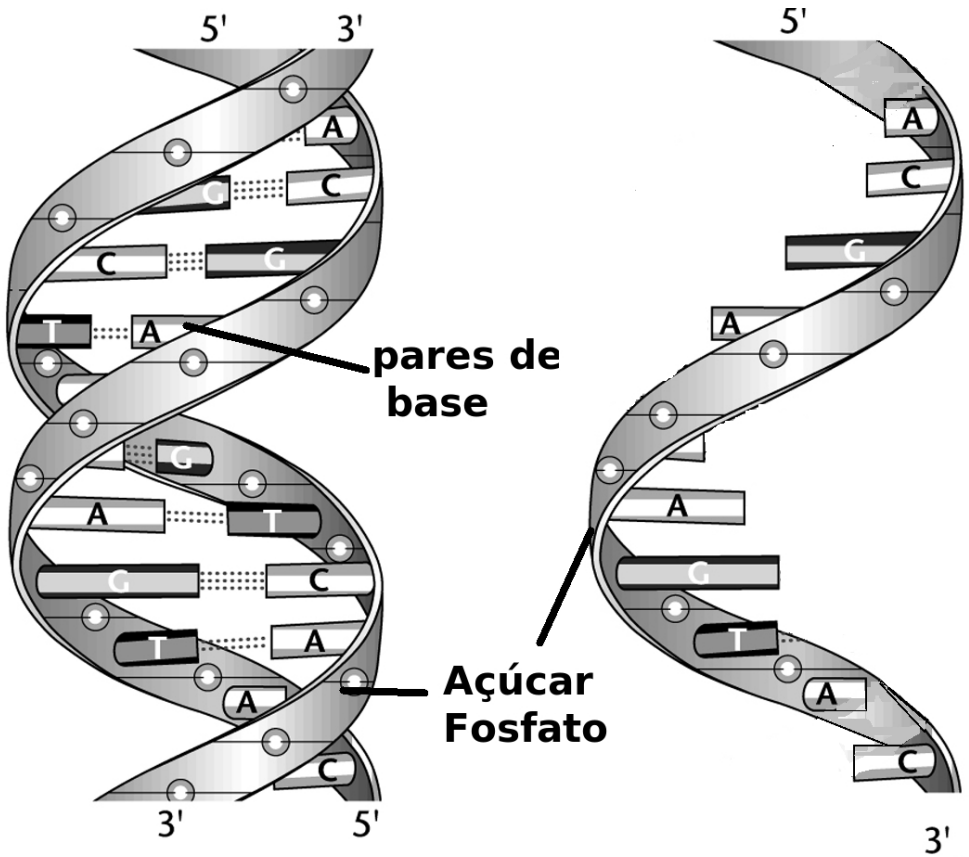


Figura: Estrutura do DNA e RNA. [**?**, Adaptada]

Transcrição

- A transcrição consiste na formação do RNA a partir do DNA.
- É feita uma cópia exata de um segmento de DNA.
- Parte do RNA formado será usado na síntese de proteínas.
- Todo esse processo é conhecido como dogma central.

Transcrição

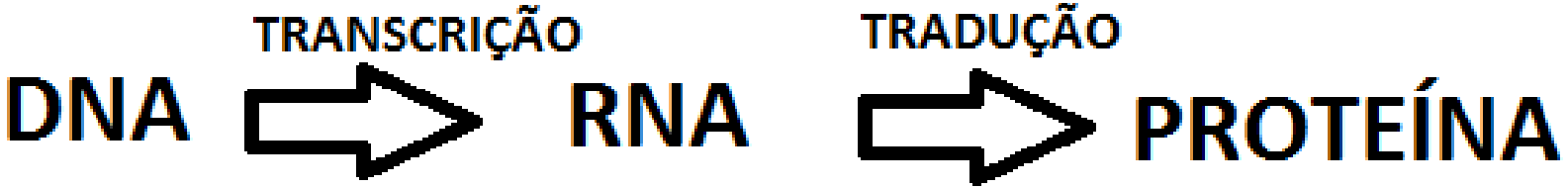


Figura: Principais passos da expressão de genética

Transcrição

- Para que ocorra a transcrição é necessário a ação de uma enzima chamada RNA-polimerase.
- Ela se conecta no DNA juntamente com fatores de transcrição gerais (formando um complexo)
- A RNA-polimerase se movimenta na direção 5' $\rightarrow$ 3' formando o RNA.

Transcrição

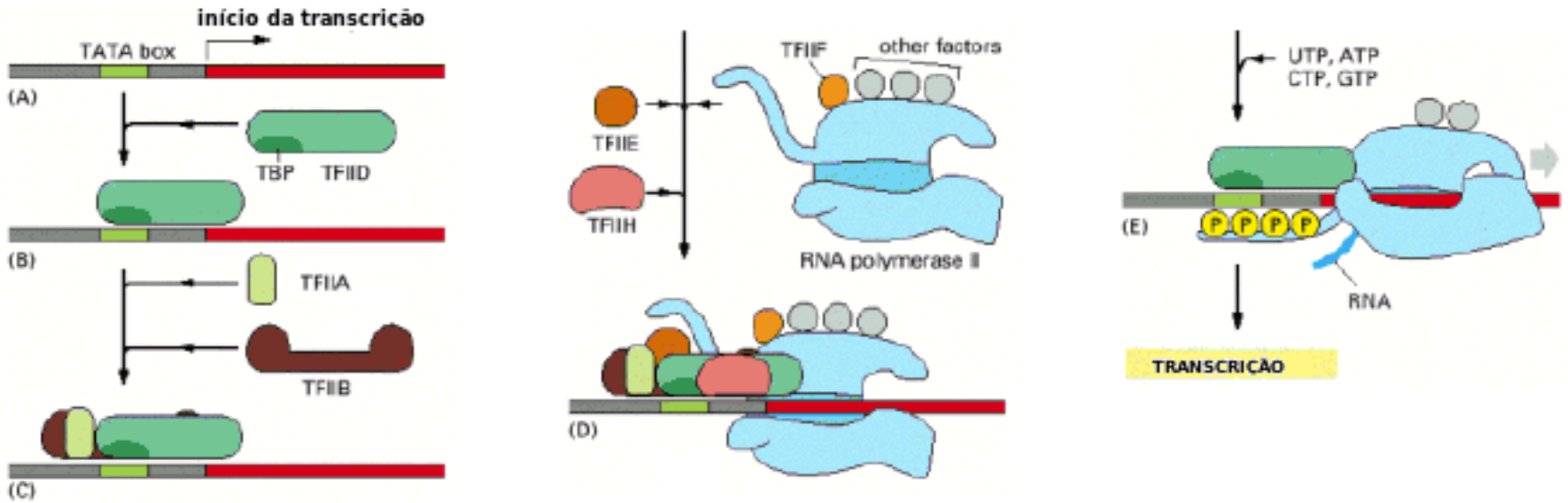


Figura: RNA polimerase e os fatores de transcrição gerais [**?**, Adaptada]

Transcrição

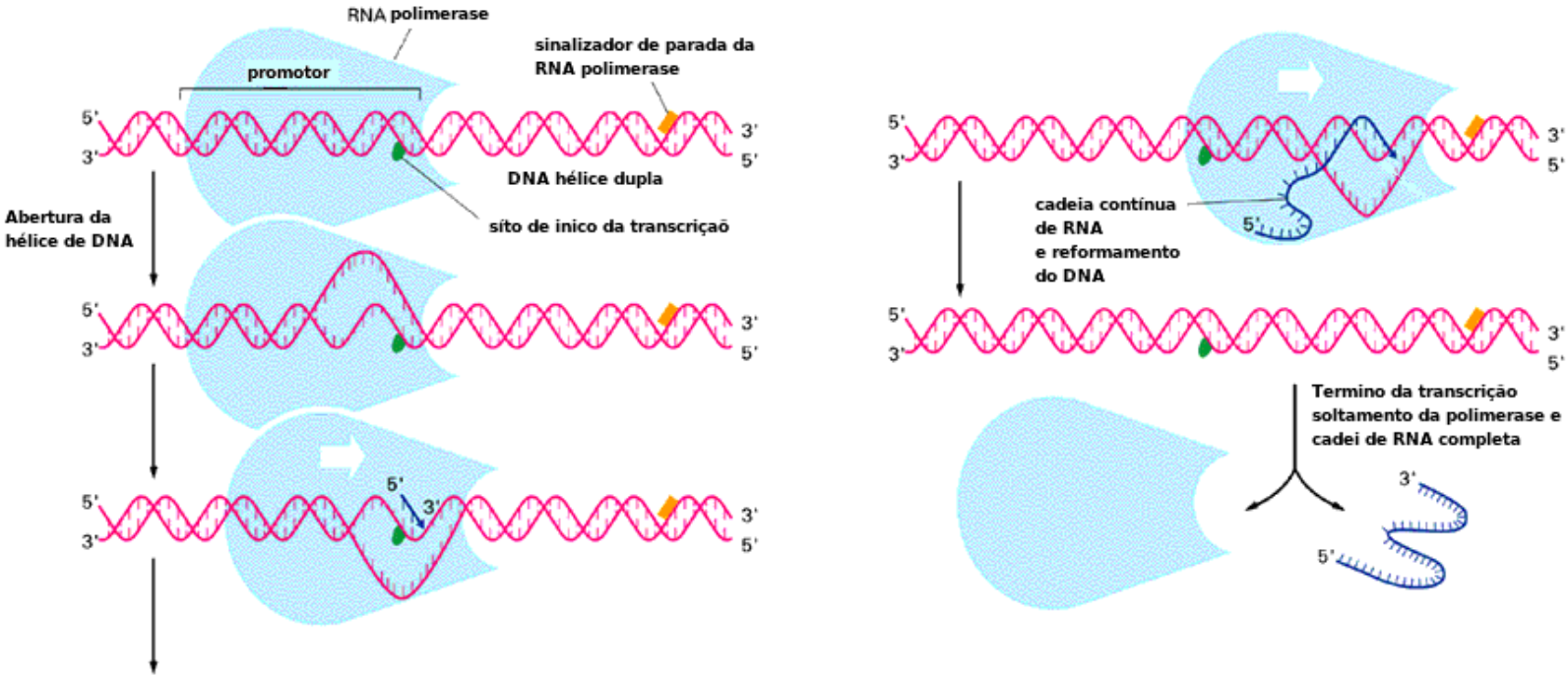


Figura: Formação do RNA através da RNA polimerase [**?**, Adaptada]

Sumário

1 Transcrição

2 Elementos regulatórios e Fatores de transcrição

3 Dehydration Responsive Binding Proteins(DREBs)

4 Busca de elementos regulatórios

5 Identificação de genes alvos de DREBs

6 Conclusão

Região reguladora

- Cada gene tem uma região regulatória, geralmente de 100-1000 pares de bases acima do local de início da transcrição.
- Dentro dela estão os elementos regulatórios.

Elementos regulatórios

- São pequenas sequências de DNA localizadas a uma distância aproximada de -50 pares de base do local de início da transcrição na região promotora de um gene (figuras 8 e 9).
- O tamanho aproximado dos elementos regulatórios varia entre 5 a 20 nucleotídeos.
- Cada elemento regulatório é específico a um fator de transcrição.

Elementos regulatórios

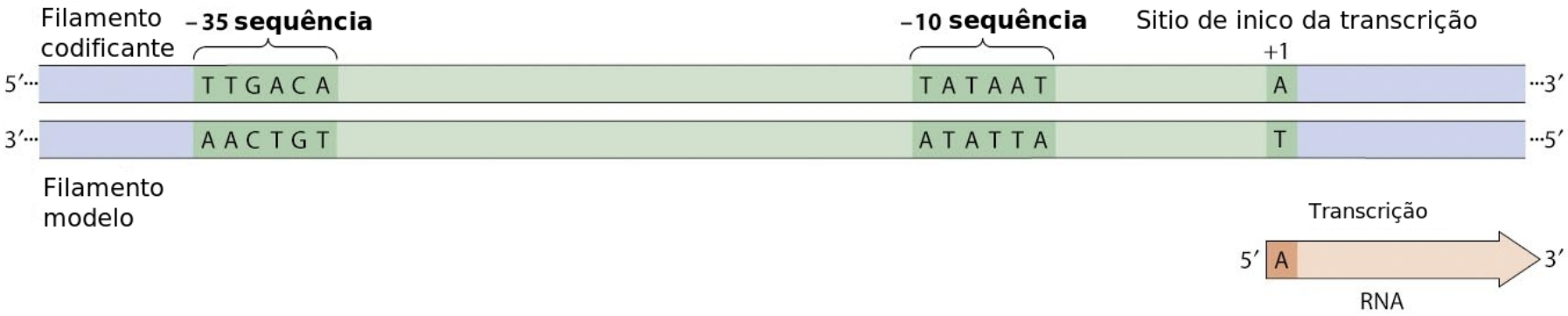


Figura: Região promotora e as sequências consenso

Elementos regulatórios

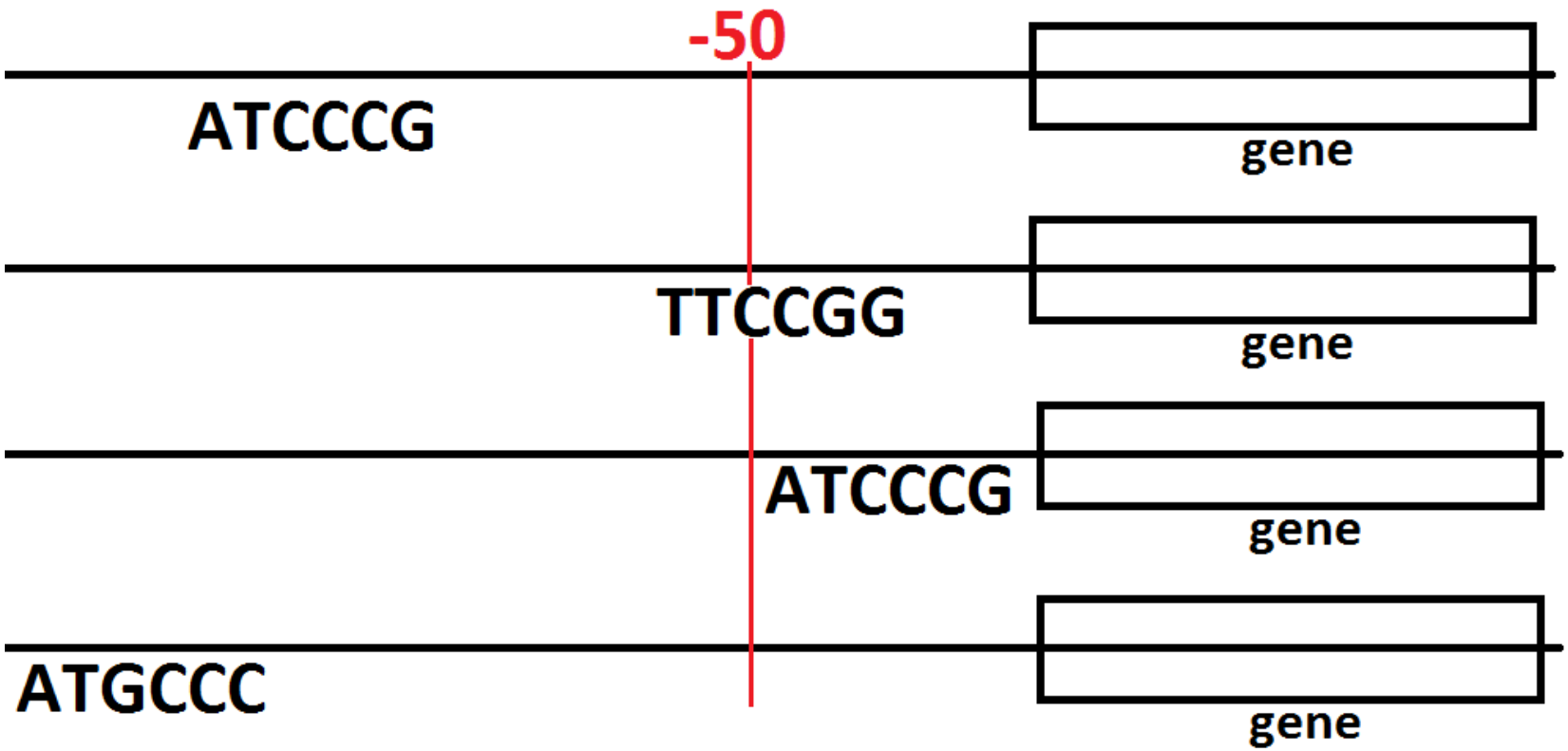


Figura: Localização aproximada dos elementos regulatórios

Fatores de Transcrição

Proteínas "especiais" chamadas de **fatores de transcrição** (TF do inglês *transcription factor binding site*) se ligam nos elementos regulatórios, contribuindo para o início da transcrição de um gene. Podem ser separados em fatores de transcrição gerais e específicos. (figuras 6 e 10).

Fatores de Transcrição

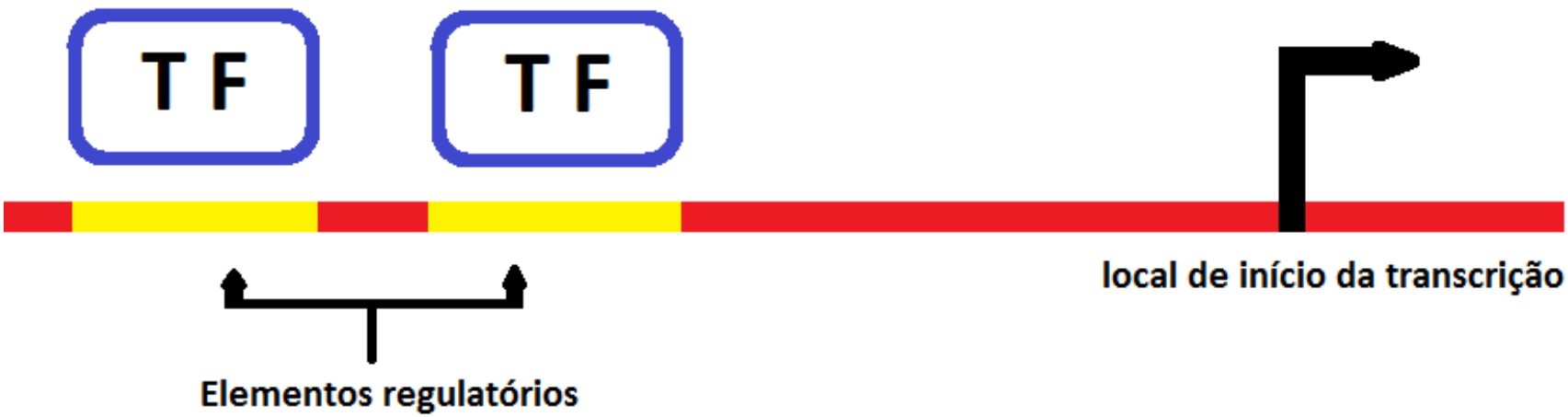


Figura: Localização aproximada dos elementos regulatórios

Funcionalidade na célula

Os elementos regulatórios juntamente com os fatores de transcrição funcionam como mecanismos de respostas a diversos estímulos, como:

- estímulos internos.
- estresses abióticos.
- estresses bióticos.

Funcionalidade na célula

Com a ativação de um elemento regulatório ocorre a expressão de um gene, que o elemento regula, o gene será transcrito no RNA que posteriormente será traduzido, gerando proteínas para suprir as necessidades do organismo.

Sumário

1 Transcrição

2 Elementos regulatórios e Fatores de transcrição

3 Dehydration Responsive Binding Proteins(DREBs)

4 Busca de elementos regulatórios

5 Identificação de genes alvos de DREBs

6 Conclusão

DREBs

São fatores de transcrição que agem na célula, quando está é estimulada por um estresse abiótico como:

- seca
- alta salinização
- baixas temperaturas.

DREBs

O entendimento dos DREBs na regulação de um gene é de grande importância para o desenvolvimento de plantas tolerantes a estresses.

Sumário

1 Transcrição

2 Elementos regulatórios e Fatores de transcrição

3 Dehydration Responsive Binding Proteins(DREBs)

4 Busca de elementos regulatórios

5 Identificação de genes alvos de DREBs

6 Conclusão

Busca por padrões

- A busca por elementos regulatórios remete a busca por padrões em uma *string*.
- Dado um padrão de DNA, encontrar sequências candidatas é simples, mas diferenciar sítios reais dos não reais é difícil.
  - Um fator de transcrição específico utilizado na expressão de um determinado gene, pode não ser o mesmo para a expressão de outro gene. Entretanto um único também TF pode regular múltiplos genes.
  - Essa especificidade torna os elementos regulatórios em sequências que não são consenso.
  - Pelo fato que muitos elementos regulatórios são usualmente degenerados, sofrem mutações, deleção ou inserção.

Um exemplo da degeneração

A

Sequence	Scores derived from matrix C (below)
CCCATTGTTCTC	18.79
TTTCTGGTTCTC	10.46
TCAATTGTTTAG	10.82
CTCATTGTTGTC	14.41
TCCATTGTTCTC	18.00
CCTATTGTTCTC	17.11
TCCATTGTTCGT	13.60
CCAATTGTTTTG	13.87

B

Position:	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	2	7	0	0	0	0	0	0	1	0
C	4	6	4	1	0	0	0	0	0	5	0	5
G	0	0	0	0	0	1	8	0	0	1	1	2
T	4	2	2	0	8	7	0	8	8	2	6	1
IUPAC:	Y	Y	H	M	T	K	G	T	T	B	D	B

Figura: Oito conhecidos elementos regulatórios da *Saccharomyces cerevisiae*, a pontuação esta de acordo com a PWM em C

Um exemplo da degeneração

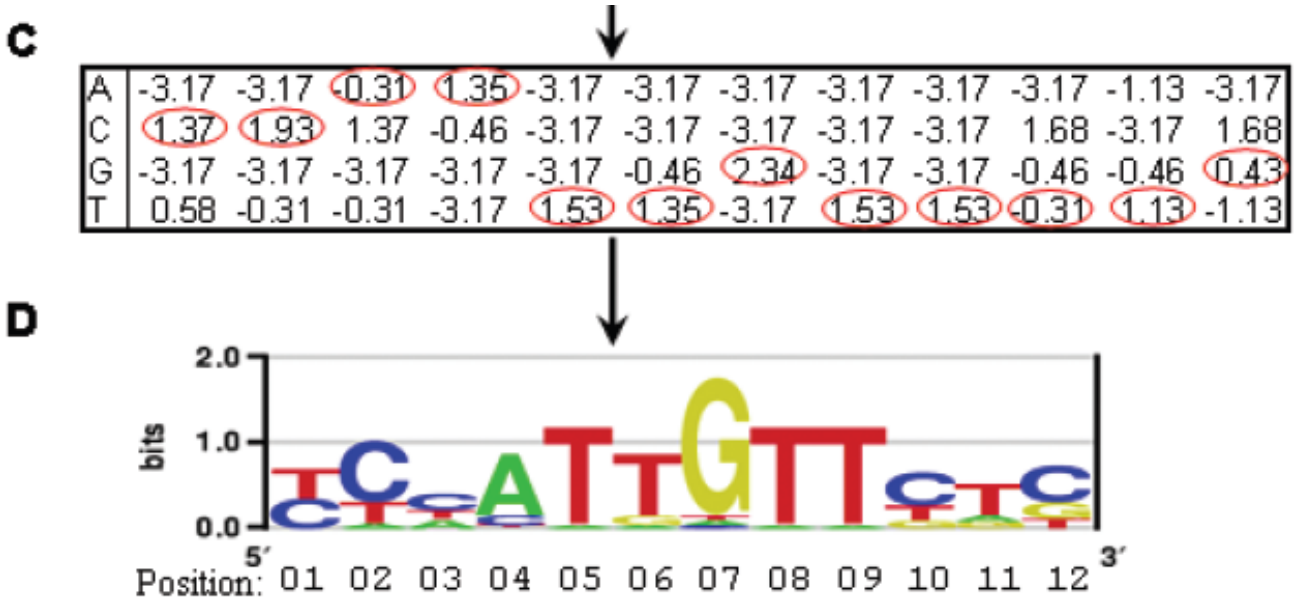


Figura: Matriz de peso das sequências alinhadas e a representação logo da sequência

Um exemplo da degeneração

Os valores da matriz de peso mostrada em C são obtidos calculando: $\log_2(f_{i,j}/P_i)$, onde $f_{i,j}$ é a frequência da base i na posição j. Este valor pode ser obtido dividindo o valor da célula da matriz pelo numero de sítios na posição (p.e $f_{C,1} = f_{T,1} = 4/8 = 0.5$). E P_i é a probabilidade de encontrar uma base na posição i, que é $P_A = P_T = 0.32$, e $P_C = P_G = 0.18$ (valor correspondente ao genoma do *S.cerevisiae*).

Busca de elementos regulatórios

- Para contornar esses problemas foram desenvolvidos vários algoritmos de busca de elementos regulatórios. Eles são classificados em três grupos:
  1. busca de em genes co-regulados
  2. busca em genes ortólogos
  3. busca simultânea em genes ortólogos e co-regulados.

Busca em genes co-regulados

- Genes que são regulados pelos mesmos conjuntos de fatores de transcrição.

Busca em genes co-regulados

seq 1	atgaccgggatactgatagaagaaaggttgggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
seq 2	acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacaataaaacggcggga
seq 3	tgagtatccctgggatgacttaaaataatggagtggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga
seq 4	gctgagaattggatgcaaaaaaagggattgtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
seq 5	tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatataataaaggaagggcttatag
seq 6	gtcaatcatgttcttgtgaatggatttaacaataagggctgggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
seq 7	cggttttggcccttgttagaggcccccgtataaacaaggagggccaattatgagagagctaatctatcgcgtgcgtgttcat
seq 8	aacttgagttaaaaaatagggagccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
seq 9	ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatactaaaaaggagcggaccgaaagggaag
seq 10	ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttactaaaaaggagcgga

Figura: Diferentes sequências promotoras de uma mesma espécie

Busca em genes co-regulados

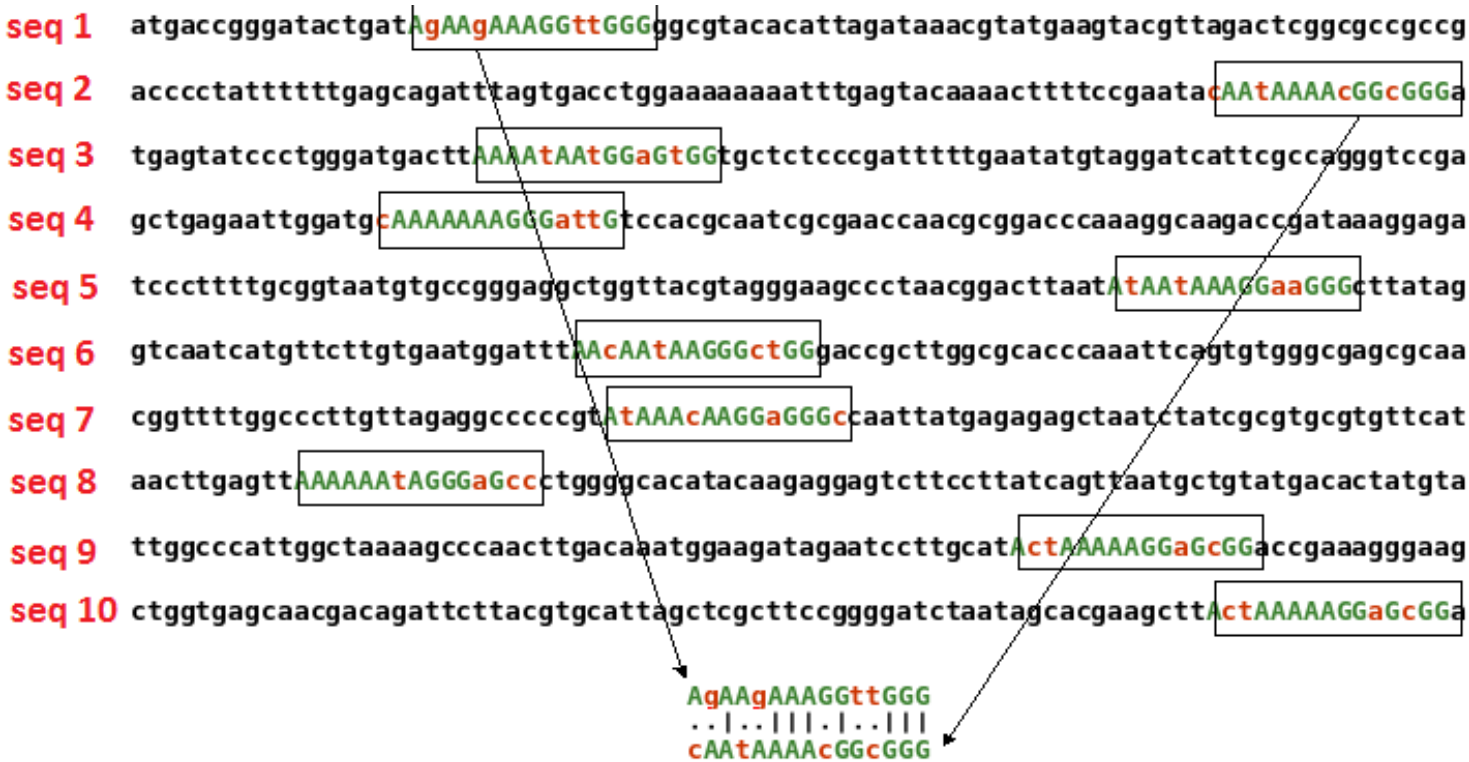


Figura: Encontrado um padrão nas sequências

Oligo-Analysis

Um dos métodos propostos para a identificação dos elementos regulatórios em genes co-regulados é o de [**?**]. Os autores projetaram um algoritmo que detecta oligonucleotídios (um fragmento curto de DNA) sobre-representados na região promotora dentro de um grupo de genes co-regulados. O programa conta todas as ocorrências dos oligonucleotídios dentro do conjunto de sequências e estima a significância estatística.

Oligo-Analysis

Primeiramente foram calculadas as frequências esperadas $F_e\{b\}$ para cada oligonucleotídio (b) de tamanhos de um a nove. Então determinada a ocorrência esperada para cada oligonucleotídio no conjunto de sequências promotoras com a formula $E(occ\{b\}) = F_e\{b\} * T$ onde $T = 2xSx(L - w + 1)$, onde w é a tamanho do oligonucleotídio; S é o numero de sequências no conjunto; L é o tamanho das sequências. O fator 2 é devido a soma de ocorrências é feita em ambos os filamentos de DNA.

Oligo-Analysis

A significancia estatística é encontrada através da formula $P(ooc\{b\} = n) = \frac{T!}{(T-n)!xn!} x(F_e\{b\})^n x(1 - F_e\{b\})^(T - n)$.

Oligo-Analysis

Depois que são encontrados os oligonucleotídios que são sobre-representados (que seguem um padrão) que têm grandes possibilidades de aparecerem em sequências promotoras, então são determinados nas sequências promotoras as posições que batem com os oligonucleotídios encontrados.

Oligo-Analysis

Este tipo de técnica garante bons resultados, até mesmo em padrões degenerados. Porém a procura por padrões com tamanhos grandes em um espaço de 4^L, onde L é o tamanho da sequência,tem um grande custo computacional, fazendo que se torne inviável para buscas com $L > 10$.

Busca em genes ortólogos

- São genes *homólogos* que foram separadas por um evento especial, fazendo que diferentes espécies tenham os mesmos genes.

Busca em genes ortólogos

```
>HUMAN
CAGGTTATCAGCAACAACACAGTCATATCCATTCTCAATTAGCTCTACCACAGTGTGTGAACCAATGTATCCAGCACCAC
CTGTAACCAAAACAATTTTAGAAGTACTTTCACTTTGTAACTGAGCTGTCATTTATATTGAATTTTCAAAAATTCTTACTTT
TTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACATATCCATATCTAATCTTACTT
ATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTCAGTAATACGCTTAAC
GCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGC
TCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAAT
CTAGCTTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACA
CCATAGGATGATAATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTA
>RAT
CGGTTTAGCATCATAAGCGCTTATAAATTTCTTAATTATGCTCGGGCACTTTTCGGCCAATGGTCTTGGTAATTCCTTTGCGC
TAGAATTGAACTCAGGTACAATCACTTCTTCTGAATGAGATTTAGTCATTATAGTTTTTTCTCCTTGACGTTAAAGTATAGAG
TATATTAACAATTTTTTGTTGATACTTTTATGACATTTGAATAAGAAGTAATACAAACTGAAAATGTTGAAAGTATTAGTTAAA
>MOUSE
CATTAATTTTGCTTCCAAGACGACAGTAATATGTCTCCTACAATACCAGTTTCGCTGCAGAAGGCACATCTATTACATTTACTG
AGCATAACGGGCTGTACTAATCCAAGGAGGTTTACGGACCAGGGGAACTTTCCAGATTCAGATCACAGCAATATAGGACTAG
```

Figura: Sequências de várias espécies

Busca em genes ortólogos

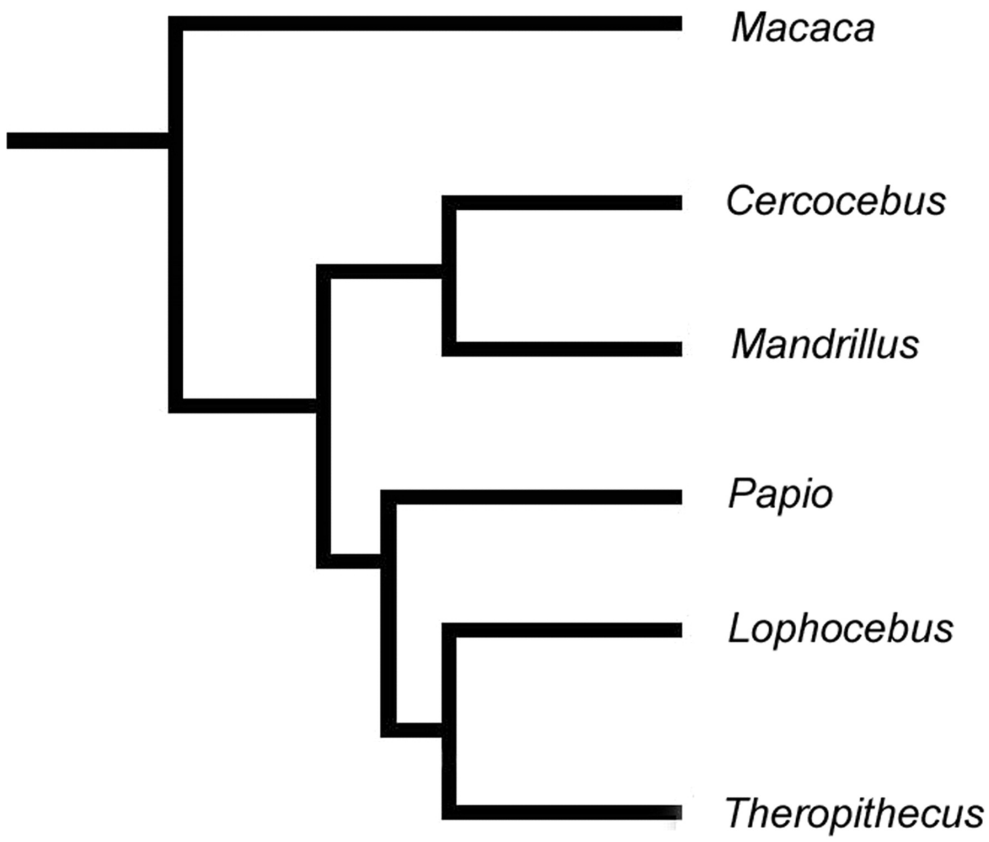


Figura: Árvore filogenética

FootPrinter

- Este algoritmo [**?**] tem como entrada uma a árvore filogenética e as sequências s promotoras de várias espécies e o tamanho do elemento regulatório.
- É calculado a menor pontuação de parcimônia da sub-árvore $d_v^* = \sum_{w \in C(v)} min_{t' \in \sum^k} (d_w^*(t') + d(t', t)$ até a raiz.
- Então é gerado sequências randômicas r que simulam a evolução das sequências s mas sem pressão natural.
- Calula-se a divergência entre s e r.
- s e r têm a mesma frequência de nucleotídeos.

Sumário

1 Transcrição

2 Elementos regulatórios e Fatores de transcrição

3 Dehydration Responsive Binding Proteins(DREBs)

4 Busca de elementos regulatórios

5 Identificação de genes alvos de DREBs

6 Conclusão

Identificação de genes alvos de DREBs

Para identificar os genes alvos de DREBs [**?**] criaram a seguinte estratégia:

Identificação de genes alvos de DREBs

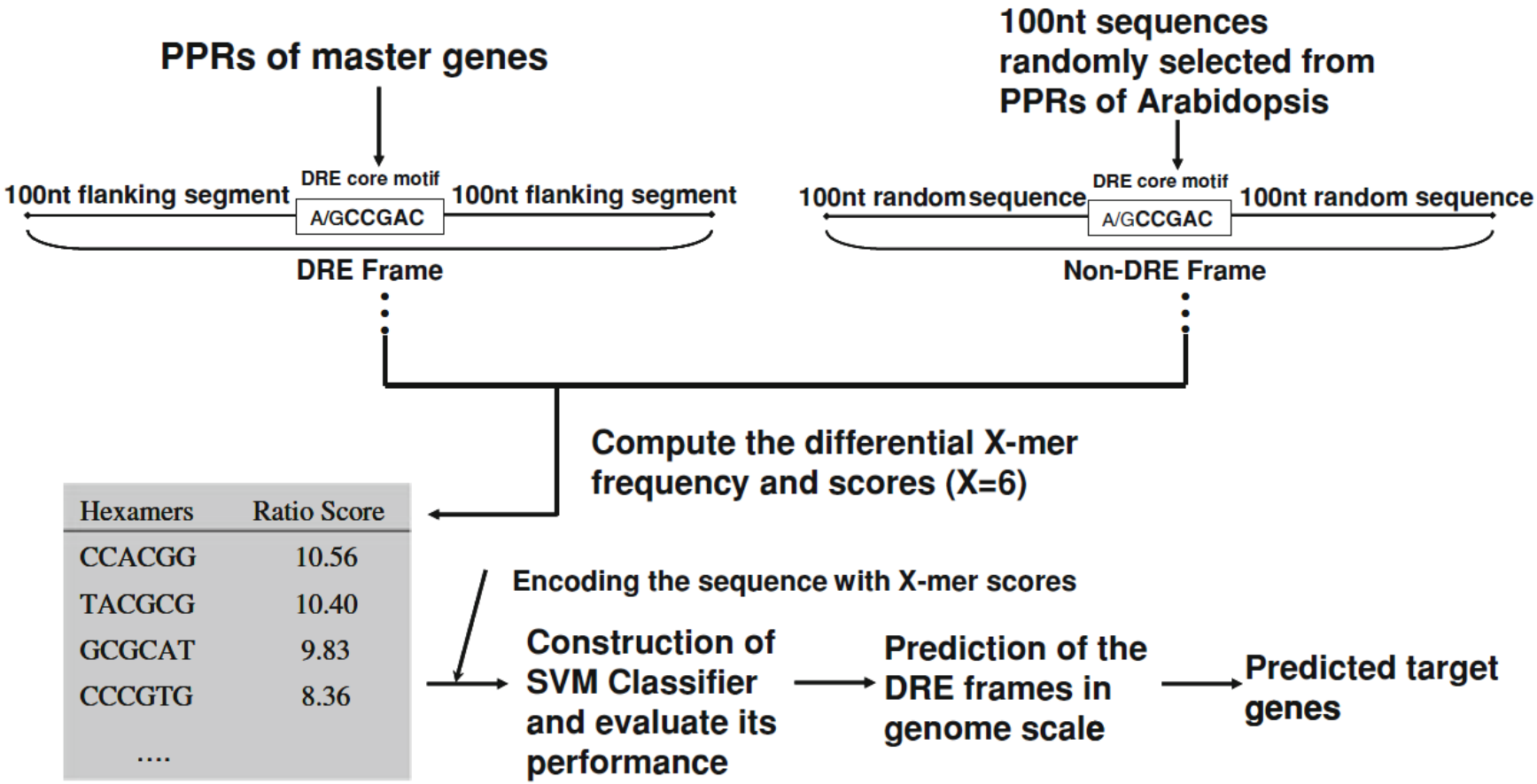


Figura: Busca DREBs

Identificação de genes alvos de DREBs

- Achar a frequência em ambos os conjuntos $Fp\{h\} e F_n\{h\}$.
- Calcular a razão $R\{h\} = \frac{Fp\{h\}}{F_n\{h\}}$
- Os DFS e nDFS recebem identificação de (+1) e (-1), respectivamente.
- SVM é treinada para distinguir entre DFS e nDFS.

Sumário

1 Transcrição

2 Elementos regulatórios e Fatores de transcrição

3 Dehydration Responsive Binding Proteins(DREBs)

4 Busca de elementos regulatórios

5 Identificação de genes alvos de DREBs

6 Conclusão

Conclusão

Dos métodos apresentados para encontrar elementos regulatórios, em específico na soja seria mais adequado a utilização de métodos de genes co-regulados. Quanto aos DREBs, utilizar uma abordagem "reversa"(com os elementos regulatórios encontrar os genes alvos), também apresenta-se mais viável do ponto de vista da utilização da ferramenta.

Bibliografia