Perfis de flexibilidade de regiões promotoras de organismos eucariotos

Dissertação de mestrado

DENISE FAGUNDES LIMA
GERALD WEBER (DF/UFMG, ORIENTADOR)

Núcleo de Pesquisas em Ciências Biológicas (NUPEB)
Pós graduação em Biotecnologia
Área de concentração: Genômica e Proteômica
Universidade Federal de Ouro Preto
Ouro Preto, maio de 2012

Perfis de flexibilidade de regiões promotoras de organismos eucariotos

DENISE FAGUNDES LIMA
GERALD WEBER (DF/UFMG, ORIENTADOR)

DISSERTAÇÃO DE MESTRADO UNIVERSIDADE FEDERAL DE OURO PRETO COMO PARTE DOS REQUISITOS BÁSICOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM BIOTECNOLOGIA, ÁREA DE CONCENTRAÇÃO GENÔMICA E PROTEÔMICA.

INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS - ICEB
UNIVERSIDADE FEDERAL DE OURO PRETO
Ouro Preto, maio de 2012

Dedicatória

Este trabalho é dedicado à minha família, especialmente aos meus pais, Neila e Luiz, e às minhas irmãs, Érica e Luiza.

Agradecimentos

De uma forma geral, eu gostaria de deixar meu agradecimento a todos que me apoiaram nesta minha trajetória. Mesmo longe, foi possível perceber a presença e o carinho de cada um de vocês. Nos tópicos abaixo estão os meus agradecimentos especiais para:

- o Dr. Professor Gerald Weber, pela orientação, pelas oportunidades e pela convivência. Agradeço muito por seu ensinamentos, carinho e, principalmente, pela paciência e confiança que teve em mim.
- o grupo de Biofísica Computacional e Física Estatística da UFOP e UFMG.
- os meus companheiros da sala 4191, Lucas, Tauanne, Míriam, Caio, Guilherme e Luciana
- os colegas do mestrado em Biotecnologia da UFOP.
- o CNPq e Capes, pela bolsa.
- o Departamento de Física da UFMG, pela acolhimento.
- a Pró-reitoria de Pós-Graduação (PROPP/UFOP), pelo auxílio dado às participações em congressos e conferências.
- os meus pais, pela educação que me forneceram, pela confiança que depositaram em mim e pelo amor incondicional.
- as minhas irmãs Érica e Luiza, pela companhia e pelo carinho.
- a minha querida amiga Luciana, pela companhia, por me ouvir e dividir anseios e alegrias durante estes dois anos de trabalho.
- todos meus amigos queridos, em especial Érica, Alice, Alexandra.
- a República Algodão Doce.
- a turma de Bacharelado em Ciências Biológicas 05/2 da UFOP.

“ I may not have gone where I intended to go, but I think I have ended up where I needed to be”

The Hitchhiker’s Guide to the Galaxy

Resumo

Em uma sequência de DNA, a ativação de um gene está relacionada com a capacidade dos fatores de transcrição (TFs) em reconhecer um região específica do DNA denominada de região promotora. Apesar da existência dos promotores *"core-less"* (conhecidas por não conter elementos regulatórios), em eucariotos esta região é caracterizada por possuir vários elementos regulatórios na porção núcleo do promotor: São eles: TATA-*box*, Iniciador (INR), elemento posterior do promotor (DPE), elemento de reconhecimento do TFIIB e as ilhas CpGs.

A busca por TSS (do inglês, *transcription start site*) e promotores é um passo importante para entender como funciona a regulação gênica e, por isso, muitos programas computacionais têm sido desenvolvidos. Além das características biológicas, alguns programas são construídos com base nos parâmetros físicos dos promotores como, por exemplo, a flexibilidade do DNA. Em geral, as técnicas experimentais utilizam sequências curtas *(n-mer)* para medir os parâmetros físicos e os valores encontrados para estas pequenas sequências são combinados para representar uma sequência maior. A maneira fisicamente correta de combinar os parâmetros de flexibilidade é através da soma inversa dos valores de elasticidade de cada *n-mer*, de forma similar ao que ocorre em uma associação de molas em série, entretanto, a maioria dos trabalhos simplesmente somam estes valores.

Diante disso, nós desenvolvemos e aplicamos um modelo de soma inversa da flexibilidade em sequências promotoras *"core-less"* e não *"core-less"* de eucariotos e construimos perfis de flexibilidade média onde avaliamos a utilização de diversos tamanhos de *n-mer*. Nós vimos que, em geral, a porção *downstream* ao TSS apresenta-se mais rígida do que a porção *usptream* ao TATA-*box* e a posição em torno de -28 é uma das regiões mais flexíveis. Nossos resultados também mostraram que a média da flexibilidade não é um boa estratégia para buscar por promotores e que, dependendo do tamanho do *n-mer* empregado, a frequência de cada valor de flexibilidade apresenta-se diferente. Ainda neste trabalho, nós analisamos a relação do conteúdo CG dos genes de microRNAs (miRNAs) e mirtrons de invertebrados e vertebrados e comparamos com a suas vizinhanças genômicas na tentativa de compreender a biogênese dos mirtrons. Mirtrons são um tipo especial de miRNA que se originam do mecanismo de *splicing* do pri-miRNA em vez de ser processado pela Drosha (enzima que cliva na haste do *hairpin* do pri-miRNA). Nós descobrimos que os mirtrons de invertebrados, até agora sem exceção, têm menor CG conteúdo do que suas regiões vizinhas. A partir deste resultado, nós sugerimos que o *splicing* que ocorre na biogênese dos mirtrons se deve ao mesmo motivo do *splicing* que ocorre durante a processamento do mRNA.

Abstract

In a DNA sequence, the gene activation is related to the ability of transcription factors (TFs) to recognize a particular region of DNA called promoter region. Despite the existence of *"core-less"* promoters (known to not contain regulatory elements), in eukaryotes this region is characterized by containing multiple regulatory elements in the portion of the core promoter: TATA-*box* Initiator (INR), downstream core promoter element (DPE), TFIIB recognition element and the CpG islands.

The search for TSS (transcription start site) and promoters is an important step to understand how gene regulation works and, therefore, many computer programs have been developed. In addition to the biological characteristics, some programs are built based on physical parameters of promoters such as the DNA flexibility. In general, experimental techniques use short sequences *(n-mer)* for measuring physical parameters and the values for these short sequences are combined to represent larger sequences. The correct way of combining the physical parameters of flexibility is by inverse sum of the values of elasticity of each *n-mer*, similarly to what occurs in a combination of springs in series. However, in most authors simply add these values together.

Therefore, we developed and applied a model of inverse sum of flexibility in *"core-less"* and not *"core-less"* promoters sequences of eukaryotes and we build profiles of average flexibility where we evaluated the use of several lengths of *n-mer*. We observed that, in general, the downstream portion to TSS is more rigid than the portion usptream to the TATA-*box*, and the position around -28 is one of the most flexible. Our results also showed that the average flexibility is not a good strategy to search for promoters and that, depending on the length of the *n-mer* used, the frequency of each value of flexibility is very different. In this work we also analyze the relationship between GC content of microRNAs (miRNAs) and mirtrons genes of invertebrates and vertebrates and compared with their genomic neighborhood in an attempt to understand the biogenesis mirtrons. Mirtrons are a special type of miRNA originating from the mechanism of *splicing* of the pri-miRNA instead of being processed by Drosha (enzyme that cleaves the stem of the hairpin the pri-miRNA). We found that mirtrons invertebrates, so far without exception, have a lower GC content than their surrounding regions. From this result, we suggest that the *splicing* that occurs in the biogenesis of mirtrons have the same reason to the *splicing* which occurs during processing of the mRNA.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O principal experimento que consolidou a genética como ciência foi o realizado por Gregor Mendel, em 1 860. Mendel descobriu que a herança das características genéticas era transmitida, de geração em geração, por meio do que ele chamou de "fatores" [1]. Atualmente, graças aos avanços da genética, sabemos que os "fatores" a que ele se referia é o que, hoje, conhecemos como genes. Naquela época, gene era a unidade molecular responsável pela hereditariedade de qualquer organismo enquanto que, hoje, genes são sequências específicas de DNA que contêm todas as instruções necessárias para coordenar o desenvolvimento e o funcionamento de todos os organismos, através da produção de proteínas e RNA funcionais. Estas instruções, porém, não são transferidas diretamente do DNA para as proteínas, o que torna necessária a transcrição do DNA em uma molécula intermediária, denominada RNA mensageiro (mRNA) [2, 3].

Tanto o DNA quanto o RNA são moléculas de natureza ácida, ricas em fósforo e nitrogênio, conhecidas, respectivamente, como ácido desoxirribonucléico e ácido ribonucléico. O nome destas moléculas se deve a presença de um tipo de açúcar com 5 átomos de hidrogêncio em cada uma delas: a desoxirribosse, no caso do DNA e, a ribose, no caso do RNA. Estudos envolvendo a degradação destes ácidos revelaram que o nitrogênio presente nestas moléculas são provenientes de bases aminadas cíclicas dos grupos das purinas e das pirimidinas. No DNA, as bases nitrogenadas púricas são representadas pela adenina e guanina e as bases pirimídicas correspondem a citosina e a timina. O RNA também possui a adenina e a guanina como bases púricas porém, suas bases pirimídicas são representadas pela citosina e a uracila [4].

Em meados do século 20, o modelo proposto para retratar a estrutura do DNA revelou que ele é formado por duas cadeias de nucleotídeos com polaridades opostas, dispostas em hélice ao redor de um eixo imaginário girando para a direita (Figura 1). Os nucleotídeos são as unidades de repetição, constituídos por uma base nitrogenada ligada a um açúcar e a um fosfato e as cadeias interagem por meio de pontes de hidrogênio que se estabelecem entre os pares de bases específicos: adenina com timina e citosina com guanina [5].

Apesar de todas as células carregarem a mesma informação genética, os genes são expressos de maneiras diferentes em cada uma delas devido ao fato de estarem submetidos a um processo muito complexo, conhecido como regulação da expressão gênica. O principal objetivo deste processo é definir o local, a quantidade e o tempo de aparecimento dos produtos funcionais de cada gene, promovendo a diferenciação celular, a morfogênese e a adaptabilidade dos organismos nos mais diversos ambientes. Por ser a etapa onde se inicia o fluxo das informações biológicas, a transcrição é a etapa onde a regulação da

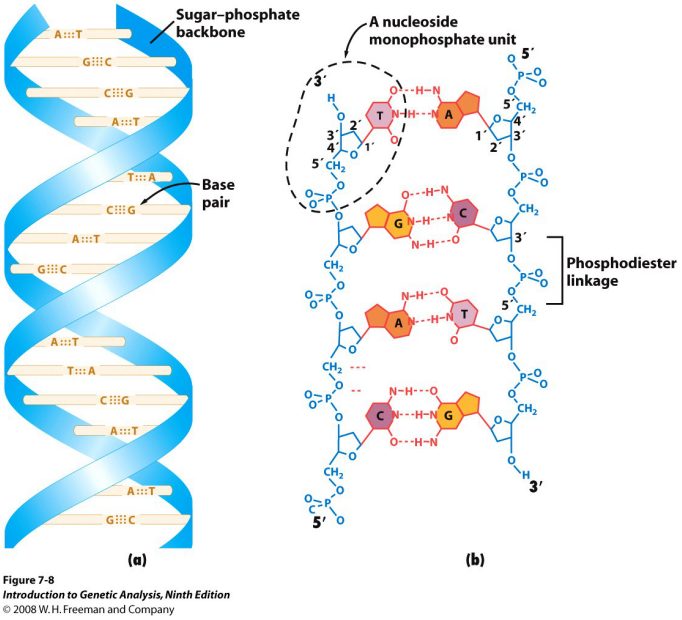


Figura 1
Representação da dupla hélice do DNA: (a) As duas cadeias de nucleotídeos são mantidas por ligações de hidrogênio existentes entre as bases nitrogenadas de cada cadeia, formando uma estrutura em espiral. (b) Cada cadeia de nucleotídeo é formada por unidades alternadas de açúcar e fosfato, unidas por ligação fosfodiéster. [1, 5]. Figura retirada do livro [1]

expressão gênica apresenta-se mais crítica. Durante a transcrição, a molécula de DNA abre-se em um determinado ponto que contém um gene e os nucleotídeos livres na célula vão se pareando a esse segmento aberto. Completado o pareamento, a molécula do RNA está pronta e o DNA, que serviu de molde, reconstitui a molécula original [1, 6, 7].

Entender como a regulação da expressão gênica atua durante a transcrição não é uma tarefa fácil. Muitas informações já foram alcançadas a respeito deste processo, porém, sabemos que ainda há muito para se descobrir. Quanto mais genes são descobertos, mais dados são disponibilizados para tentar compreender a regulação da expressão gênica e, por conta disso, estudos envolvendo a identificação gênica vêm assumindo papel de destaque no ramo das pesquisas. A maioria dos genes conhecidos e anotados nos principais bancos de dados foram detectados através de uma técnica laboratorial que sintetiza cDNA (DNA complementar) a partir da transcrição inversa de moléculas de RNAm funcional, possibilitando a determinação direta da porção do DNA que expressa uma proteína [8, 9]. Com o cresimento das tecnologias de alta produtividade, vários genomas tiveram o seu sequenciamento completado e o grande volume de dados gerados levou a necessidade de atrelar a identificação gênica à biologia computacional, por meio dos algoritmos de predição. É importante ressaltar que os algoritmos de predição gênica não substituem as técnicas convencionais utilizadas para identificar e caracterizar as regiões promoto-

ras mas, representam uma ferramenta adicional que surgiu para atender os desafios da era genômica. Existem três maneiras para construir um algoritmo de reconhecimento de promotor: por meio das características de sinal, contexto e estrutura. No caso do nosso trabalho, nós avaliamos como o método baseado na flexibilidade de regiões promotoras têm sido empregado nesta área de pesquisa. Antes de focarmos em nossa metodologia, faremos uma breve revisão sobre os aspectos biológicos e físicos da transcrição em eucariotos, uma vez que, muitas destas informações são a base para o desenvolvimento dos algoritmos de predição [10–12].

2 Os aspectos biológicos e físicos da transcrição

2.1 A biologia da transcrição

A maioria dos genomas de eucariotos possuem dezenas de milhares de genes a serem transcritos que, em média, apresentam-se bem espaçados uns dos outros. Para identificá-los e dar início ao processo transcricional, um eficiente aparato enzimático é formado para reconhecer uma região específica dos genes, denominada região promotora [1]. A região promotra, ou simplesmente promotor, se localiza nas adjacências do sítio de início de transcrição (TSS-*transcription start site*) e é tipicamente dividida em três porções: núcleo do promotor, promotor proximal e promotor distal, sendo que o que diferencia uma porção das outras é o tipo de elemento regulatório que cada uma possui e a distância a que se encontram do TSS. Proteínas reguladoras, conhecidas como fatores de transcrição (TFs), são os principais componentes do aparato enzimático que, junto com as polimerases, identificam e se ligam nestas porções, principalmente no núcleo do promotor. O núcleo do promotor, representado pela figura 2, é a porção mais próxima do TSS, situada em torno de 40 pares de bases (pb) acima (*upstream*) e abaixo (*downstream*) do ponto inicial da transcrição e, além disso, é a região que possui os elementos regulatórios mais estudados [13, 14].

Os organismos eucariotos possuem três polimerases diferentes para reconhecer três subconjuntos de genes: pol I, pol II e pol III, veja a tabela 1. A polimerase responsável pela transcrição dos genes que codificam proteínas, pol II, se associa com os fatores de transcrição TFIIA, TFIIB, TFIID, TFIIE, TFIIF e TFIIG e formam o PIC, um complexo de pré-iniciação [12, 15–17]. Estudos com regiões promotoras de eucariotos revelaram

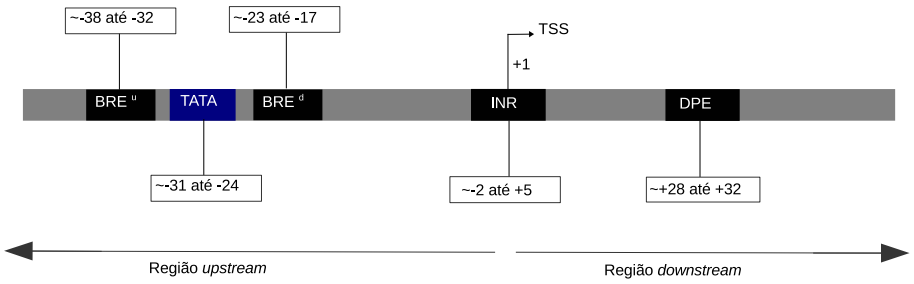


Figura 2
Representação esquemática do núcleo do promotor e dos elementos regulatórios TATA-box, iniciador (INR), elemento posterior ao promotor (DPE) e elemento de reconhecimento do TFIIB (BRE), em suas respectivas posições. Obs: As ilhas CpGs não foram representadas neste esquema pois não situam-se no núcleo do promotor.

RNA pol	transcreve
I	rRNA (exceto o 5S rRNA)
II	todos os genes que codificam proteínas presentes no mRNA e alguns snRNAs
III	pequenos RNAs funcionais como tRNA, alguns snRNA e 5S RNA

Tabela 1
Os três diferentes tipos de RNA polimerases: I, II e III.

que a existência de certos elementos regulatórios, principalmente no núcleo do promotor, otimizam a formação deste complexo, tornando a transcrição mais eficiente. Entre os elementos regulatórios mais estudados estão o TATA-box, o iniciador (INR), o elemento posterior ao promotor (DPE), o elemento de reconhecimento do TFIIB (BRE) e as ilhas CpGs (Fig. 2). Cada um deles pode ser encontrado apenas em um conjunto de promotores e, em uma mesma região promotora podem existir diferentes combinações destes elementos. Eles são representados por sequências consensos de nucleotídeos que variam em torno de 6 a 10 pb. Isso quer dizer que cada base de uma sequência consenso representa o nucleotídeo mais frequente que aparece naquela posição, sendo possível encontrar uma variedade de sequências similares às sequências consensos [18–20]. Para exemplificar, vamos considerar que as sequências BANANA, BATATA, BATANA e BANATA sejam sequências similares. A sequência consenso que melhor representa todas estas sequências é BA[T/N]A[T/N]A e os colchetes presentes nas posições 3 e 5 indicam que as letras T e N são as letras que ocorrem com maior frequência nestas posições.

O elemento TATA-*box* consiste em um elemento rico em A/T localizado, aproximadamente, 25 a 30 pb acima do TSS dos genes que codificam proteínas (Fig. 2). Ele foi o primeiro elemento a ser identificado e sua sequência consenso, TATA[A/T]A[A/T], é reconhecida por uma subunidade do TFIID, denominada proteína de ligação ao TATA (TBP) [14, 21, 22]. Apesar de ser uma sequência relativamente conservada, é importante ressaltar que nem todos promotores apresentam o TATA-*box* [16]. Em 2000, por exemplo, Kutach e Kadonaga [23] estimaram que 43% das 205 regiões promotoras de *Drosophila*, descobertas até aquele ano, apresentavam este elemento. Já em humanos, até o ano de 2001, apenas 32% dos 1031 promotores continham este elemento rico em A/T em suas sequências [13]. Dados mais recentes mostram que o TATA-*box* representa uma minoria entre os promotores de mamíferos, estando presente somente em 10%-16% dos promotores de ratos e humanos [24]. Em geral, os promotores contendo o TATA-*box* estão relacionados com a expressão de genes de tecidos específicos [25] e sua presença é comumente associada ao elemento iniciador [20].

O iniciador é um elemento rico em pirimidinas que, em mamíferos, é definido pela

sequência consenso [C/T][C/T]AN[T/A][C/T][C/T], onde N representa qualquer nucleotídeo. O inciador se localiza entre as posições -2 e +5 do TSS e, em geral, o nucleotídeo A contido em sua sequência consenso ocupa posição +1 do TSS (Fig. 2). Ele é reconhecido por duas subunidades do TFIID, TAF1 e TAF2, denominadas fatores associados ao TBP. Conforme dito no parágrafo anterior, o iniciador pode se associar ao TATA-*box* e, se a distância entre os dois elementos for de 25 a 30 pb, ambos funcionam de forma sinérgica [26]. Caso a distância seja superior a 30 pb, cada elemento irá atuar independentemente. Em regiões promotoras desprovidas de TATA-*box*, conhecidas como TATA-*less*, é comum encontrarmos o elemento iniciador e sua presença relaciona-se com a expressão de genes constitutivos, oncogenes, fatores de crescimento e fatores de transcrição [14, 22, 24, 27].

O elemento posterior ao promotor (DPE) é identificado, predominantemente, em regiões promotoras TATA-*less* de *Drosophila*, humanos e outros organismos. Está localizado entre as posições 28 e 32 abaixo do TSS (Fig. 2) e sua sequência consenso é representada por [A/G]G[A/T][C/T][A/C]. Em geral, os promotores que possuem o elemento DPE requerem a presença do INR para exercer sua função e o espaçamento entre eles é fundamental para que o complexo PIC possa reconhecê-los [14,27,28]. Subunidades de TFIID, TAF6 e TAF9, são os principais fatores que se ligam à estes elementos e uma mutação em qualquer um deles resulta na perda da atividade transcricional [22, 29]. No estudo citado anteriormente, Kutach e Kadonaga [23] também mostraram que a DPE é quase tão comum quanto o TATA-*box*. Eles estimaram que 29% dos 205 promotores de *Drosophila*, continham apenas TATA-*box* (sem DPE), 26% possuiam apenas DPE (sem TATA), 14% continham tanto DPE quanto TATA-*box* e 31% aparentemente não continham nenhum dos dois elementos. Ou seja, a proporção de TATA-*box* e DPE nas sequências promotoras são muito similares [23, 30].

O elemento de reconhecimento ao TFIIB (BRE) é o único elemento bem caracterizado do núcleo do promotor que não é reconhecido pelo TFIID, podendo ser encontrado tanto acima da região TATA-*box* (BREu), em torno da posição -38 a -32, quanto abaixo (BREd), na posição +23 a +17 (Fig. 2) [22]. O TFIIB é uma proteína que apresenta-se como uma hélice-volta-hélice, onde sulco maior reconhece e interage com o BREu e, o sulco menor, com o BREd [14, 27, 31]. A interação de TFIIB com o BREd, cuja sequência consenso é [G/A]T[T/G/A][T/G][G/T][T/G][T/G] parece depender da interação que TBP realiza junto à sequência TATA-*box* [32, 33], sendo verificado, em algumas sequências, o efeito negativo desta ligação sobre a transcrição [33,34]. Já no caso do BREu, representada pela sequência consenso [G/C][G/C][G/A]CGCC, esta interação pode ocorrer independente da ligação do TBP ao TATA-*box* e o seu contato com o TFIIB produz um efeito positivo sobre a transcrição [35].

O último elemento mais bem estudado, localizado fora do núcleo do promotor, são as ilhas CpGs onde o p representa a ligação fosfodiéster que ocorre entre as bases citosina e guanina [36]. Estes elementos são encontrados em vertebrados e, menos frequente em plantas, e a presença deles está relacionada com a expressão de genes constitutivos ou de tecidos específicos. Em genomas de mamíferos, o dinucleotídeo CpG possui baixa representação pois, grande parte deles, são metilados resultando na 5-metilcitosina, que serve como substrato para a DNA metiltransferase. Esta enzima promove a desaminação da 5-metilcitosina em resíduos de timina, que não são reconhecidos pelo sistema de reparo do DNA [22, 37]. Somente uma pequena porção destes dinucleotídeos é que constituem as ilhas CpGs, que variam em torno de 500 a 2 000 pb e, normalmente, elas são localizadas perto dos sítios de início de transcrição. Cerca de 50% dos promotores de genes de mamíferos estão associados com uma ou mais ilhas CpGs, e a metilação dessas regiões funcionam como reguladores do processo transcricional [24]. Ilhas CpGs não metiladas, próximas à regiões promotoras, correspondem a promotores ativos, enquanto que, ilhas CpGs metiladas dificultam o acesso dos fatores de transcrição, tornando os promotores inativos.

2.2 A física da transcrição

Os estudos dos aspectos biológicos da transcrição têm fornecido informações relevantes que ajudam a entender o funcionamento da regulação gênica durante a síntese protéica. Apesar de serem bem caracterizados dentro do contexto biológico, os elementos regulatórios não são suficientes para discriminar promotores de não-promotores [12]. O que comprova isso é a existência de uma classe de promotores, denominada *"core-less"*, onde todos os seus integrantes não possuem qualquer tipo de elemento regulatório e, mesmo assim, são reconhecidos pelo PIC [38, 39]. Além disso, os elementos regulatórios são constituídos por sequências degeneradas de apenas 6 a 10 pb, o que torna alta a probabilidade de encontrar sequências similares às sequências consensos em outras regiões do genoma [19]. Para exemplificar, Pedersen *et al.* [12] mostrou que uma sequência perfeitamente igual à sequência consenso do INR aparece, por acaso, cerca de uma vez a cada 512 pb em uma sequência aleatória. Outro estudo verificou que o TATA-box pode ser encontrado em uma média de 120 pb no conjunto de sequências não-promotoras de mamíferos [12]. Devido a estes fatores, os aspectos físicos e as propriedades estruturais do DNA vêm assumindo um papel de destaque neste ramo de pesquisa pois trazem informações que se complementam às informações biológicas, ajudando a entender como os fatores de transcrição reconhecem as sequências consensos dos elementos regulatórios e os promotores *"core-less"* [20, 38].

A dupla hélice de DNA é uma molécula altamente dinâmica, que pode tanto se curvar como se dobrar [40]. Sua estrutura tridimensional é mantida por interações químicas que ocorrem entre os nucleotídeos vizinhos e, principalmente, pelas três ligações de hidrogênio que unem a citosina e a guanina, e pelas duas ligações de hidrogênio que acontecem entre a adenina e a timina. Os ângulos formados por estas ligações garantem os movimentos rotacionais e translacionais à molécula, conferindo estabilidade, curvatura e flexibilidade ao DNA. Estas três propriedades são responsáveis pela conformação estrutural do DNA, podendo ser vistas regulando importantes processos biológicos como a replicação, a transcrição e a recombinação gênica. Nos parágrafos a seguir, falaremos um pouco sobre cada uma destas propriedades [11].

A estabilidade do DNA é uma propriedade muito correlacionada com o seu processo de denaturação. Regiões do DNA que são termodinamicamente mais instáveis, se desnaturam mais facilmente e se desenrolam, expondo ambas as fitas. Tanto a replicação quanto a transcrição necessitam que as duas fitas do DNA sejam desenroladas para que uma delas possa ser copiada [41]. Podendo ser expressa através da energia livre de Gibbs, a predição da estabilidade de um duplexo de DNA depende de sua sequência primária de nucleotídeos, uma vez que o cálculo considera a interação que cada par de base realiza com o seu próximo vizinho, à temperatura e pH definidos. A equação da energia livre de Gibbs,

$$\Delta G = \Delta H - T\Delta S, \tag{1}$$

utiliza valores tabulados de entalpia ΔH e entropia ΔS. Ela é calculada para cada um dos pares de bases e, em seguida, os valores de ΔG encontrados são somados para se obter o valor total da estabilidade da molécula ΔG_{total}, como é mostrado na figura 3 [42]. Se o valor de ΔG_{total} for negativo, as cadeias do DNA vão se unir espontaneamente e a molécula apresentará um caráter estável. Caso contrário, se o ΔG_{total} for positivo, a união de ambas as fitas do DNA dependerá de um algum fator externo e a establidade desta molécula será consideravelmente menor. Partindo deste modelo, Kanhere *et al.* [19] analisou o perfil de estabilidade média das sequências promotoras de quatro grupos de organismos distintos: vertebrados, plantas, *Escherichia coli* e *Bacillus subtilis*. Ao contrário dos eucariotos, os organismos procariotos como *E. coli* e *B. subtilis* apresentam apenas duas regiões muito conservadas, ambas situadas acima do sítio de início de transcrição, uma na posição -35 e a outra na posição -10, sendo que a distância entre estas regiões é fundamental para a interação da RNA polimerase II. O elemento da posição -35 é representado pela sequência TTGACA enquanto que o elemento da posição -10 é representado pela sequência TATAAT [43]. Em todos os grupos analisados, os valores de ΔG_{total} encontrados para a região *upstream* tendem a ser maiores do que os valores de ΔG_{total} encontrados

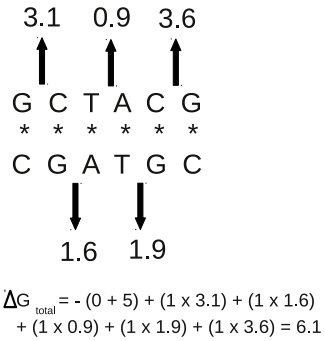


Figura 3
Exemplo do cálculo de energia livre de Gibbs: A energia livre total de um duplexo, aqui dada em kcal/mol, é a soma dos valores de energia livre de cada par de bases em relação ao seu próximo vizinho. Ref. [42] atribui 5 kcal para duplexos contendo C/G e 6 kcal para duplexos compostos exclusivamente de A/T. Além disso, se o duplexo é formado apartir de uma fita auto-complementar, é adicionado 0,4 kcal.

para a região *downstream*, levando a conclusão de que a região acima do TSS é menos estável do que a região abaixo deste sítio [19].

A curvatura do DNA é uma propriedade intrínseca ao DNA, sendo determinada apenas pelos ângulos rotacionais dos seus pares de bases [44]. Além de depender da sequência primária de bases, a curvatura do DNA também depende da temperatura e da concentração fisiológica de certos íons [45]. Sua medida pode ser realizada por meio de várias técnicas, como ressonância magnética nuclear, corrida em gel de eletroforese, cristalografia de raio-x, etc e, em geral, os dados obtidos por estas técnicas são utilizadas para predizer, computacionalmente, a curvatura de sequências genômicas, sintéticas ou não [44]. Em artigo publicado em 2005, Asayama *et.al* [46] cita alguns trabalhos que demonstram que a curvatura do DNA ocorre, frequentemente, em torno de origens de replicação, promotores, *enhancers* e sítios de recombinação, o que realça a importância desta propriedade em muitos processos genéticos básicos. Asayama *et.al* [46] enfatiza o papel da curvatura do DNA na porção do núcleo do promotor nas sequências promotoras de organismos bacterianos, mostrando que a região *usptream* ao TSS apresenta-se mais curva do que a porção *downstream* e que, nos promotores desprovidos de sequências consensos, a estrutura curva do DNA funciona como um importante sinal para a ligação do PIC ao promotor.

A flexibilidade do DNA é uma propriedade relacionada com capacidade do DNA em se dobrar ou se curvar, de forma mais acentuada, quando interage com fatores externos tal como uma proteína ou outros ligantes [44]. Dessa forma, a flexibilidade do DNA se faz presente na transcrição, na replicação e no empacotamento do material genético, onde genomas de até 3 bilhões de pb [47] passam a ocupar volumes muito reduzidos

ao interagirem com as proteínas histonas, levando à formação de unidades estruturais de repetição denominadas nucleossomos [40]. Assim como as outras propriedades, a flexibilidade também depende da composição de bases do DNA e o seu cálculo têm sido aplicado em sequências promotoras dos mais diversos organismos [20, 48]. Existem alguns modelos já estabelecidos para poder predizer a flexibilidade do DNA e a utilização deles em sequências promotoras fornecem duas conclusões que são amplamente difundidas no meio científico: a primeira é que a região *downstream* ao TSS é mais flexível do que a região *upstream* e a segunda é que o TSS e a região em torno da posição -28 são relativamente mais rígidas [49].

3 Algoritmos de predição gênica

As informações físicas e biológicas das sequências promotoras são fundamentais para os biólogos computacionais construirem os algoritmos de predição gênica. Apesar deles usarem todas estas informações, a precisão dos programas já existentes ainda não é satisfatória [50–53] e a principal razão para isso é que o nosso conhecimento atual sobre o início da transcrição não é completo. Além disso, a forma como alguns programas lidam com estas informações também contribuem para uma baixa precisão. Por exemplo, os sítios de ligação dos fatores de transcrição são funcionais em ambas as fitas de uma sequência genômica e, apesar disso, alguns algoritmos de predição gênica buscam promotores apenas em uma das fitas do DNA [50]. Existem algoritmos que são muito genéricos e, talvez, se focassem em uma determinada sub-classe de promotores, seu desempenho poderia ser melhor.

Os programas de reconhecimento de promotores são construídos, basicamente, pela escolha de uma das três características do DNA: sinal, contexto ou estrutura [11]. As características de sinal correspondem aos elementos regulatórios presentes no núcleo do promotor (TATA-*box*, INR, DPE e BRE, entre outros), aos sítios de ligação dos fatores de transcrição e às ilhas CpGs dos mamíferos. *Autogene* [51], *PromoterScan* [51, 54], *Eponine* [55] são algumas das ferramentas desenvolvidas que incorporam características de sinal em suas análises, no entanto, é importante lembrar que a simples utilização desse método não é suficiente para discriminar promotores de não-promotores. Um estudo feito com 1871 promotores de humanos depositadas no EPD (*Eukaryotic Promoter Database*) mostrou que, apenas 6, 9 e 0,4% daquele total possuem apenas o elemento TATA-*box*, apenas o elemnto INR e apenas o elemento DPE, respectivamente [38]. A baixa frequências destes elementos, aliada ao fato de serem sequências consensos de apenas 6 a 10 pb, confirmam a incapacidade deste método em reconhecer promotores de maneira eficiente [11, 19]

As características de contexto são representados por conjuntos de *n-mers* (sequências de DNA compostas por *n* bases) extraídos do contexto genômico dos promotores, os quais têm sua estatística estimada a partir de amostras de treinamento. Apesar de ser um método independente dos sinais biológicos, os *n-mers* podem, muitas vezes, envolver certos tipos de sinais tal como os TATA-*box* e ilhas CpGs, ajudando a reduzir a taxa de falsos positivos. O TATA-*box*, por exemplo, por ser representado por *6-mer*, TATA[A/T]A, e uma análise feita com as 1871 sequências promotoras de humanos depositadas no EPD mostrou que o *6-mer* mais frequente, na posição -29, corresponde exatamente ao elemento TATA-*box* [11]. Por não depender dos sinais biológicos, este método pode revelar detalhes

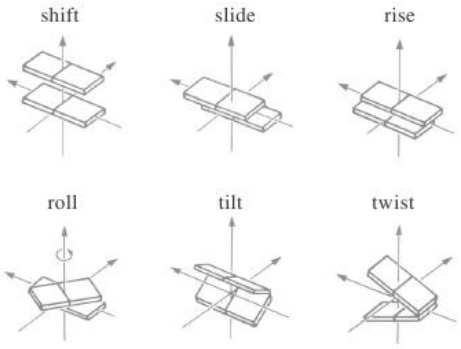


Figura 4
Esquema da movimentação dos ângulos translacionais (acima) e rotacionais (abaixo) de um par de bases e seu próximo vizinho presente na dupla hélice do DNA. Figura retirada da referência [40].

desconhecidos das regiões promotoras, o que torna interessante ser o primeiro recurso escolhido quando se pretende realizar um reconhecimento de promotor inicial em todo o genoma. Se enquadram neste método os seguintes programas: *PromFind* [56], *Promoter 2.0* [57] e *PromoterInspector* [58].

As características de estrutura são baseadas nas propriedades físicas que mantêm a conformação tridimensional do DNA. Para avaliar como estas propriedades podem ajudar no reconhecimento de promotores, são realizadas análises dos ângulos rotacionais e translacionais da dupla hélice, principalmente na porção do promotor proximal e na porção do núcleo do promotor [11]. Cada um destes ângulos, conhecidos como *shift*, *tilt*, *roll*, *slide*, *rise* e *twist*, representa um movimento específico que pode ocorrer entre dois possíveis pares de bases, como é mostrado na figura 4. Entre todas as propriedades físicas, a flexibilidade do DNA é a mais utilizada como característica de estrutura e, em geral, ela é calculada por meio de um dos quatro dados experimentais a seguir: sensibilidade à DNase I [59], posicionamento do nucleossomo [60], posição de torção da hélice [61, 62] e energia potencial de superfície [63]. Estes dados experimentais são baseados na análise de pelo menos um dos ângulos acima em *n-mers* de tamanhos reduzidos e, em função disto, a flexibilidade total de sequências maiores depende da associação de vários *n-mers*.

De acordo com a nossa revisão bibliográfica, a soma direta das flexibilidades dos *n-mers* é o tipo de associação mais utilizado para obter a flexibilidade total de sequências genômicas [11, 19, 20, 38, 49, 64–66]. Independente do modelo de flexibilidade escolhido e do tamanho do *n-mer* considerado, os trabalhos que utilizam a soma direta das flexibilidades dos *n-mers* em sequências promotoras costumam chegar a conclusões muito semelhantes. Pedersen *et al.* [48], por exemplo, construiu perfis de flexibilidade média com base na soma direta das flexibilidades dos *n-mers* para cada um dos três modelos

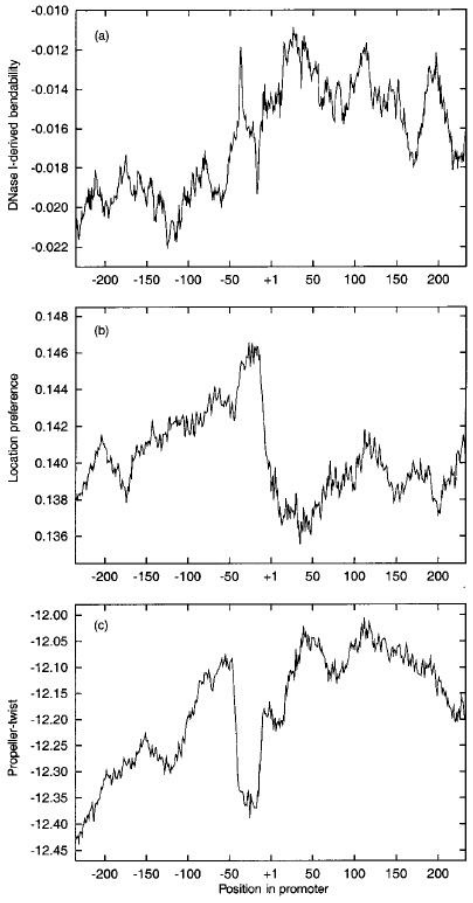


Figura 5
Exemplo de perfis de flexibilidade média das sequências promotoras de humanos depositadas no DBTSS: a posição +1 corresponde ao sítio de início da transcrição (TSS). (a) Perfil de flexibilidade média para cada posição do promotor usando o modelo derivado da sensibilidade à DNase I [59] e janela de tamanho *20- mer*. Valores altos correspondem à alta flexibilidade. Os dois picos em torno da posição -30 são causados pelo TATA-*box* presente nesta região. (b) Perfil de flexibilidade média usando o modelo de posicionamento do nucleossomo [60] e janela de tamanho *30- mer*. Valores baixos de flexibilidade correspondem à sequências mais flexíveis. (c) Perfil de flexibilidade média baseado na na posição de torção da hélice [61, 62] e janela de tamanho *30- mer*. Valores altos (menos negativos) correspondem à alta flexibilidade. Os três modelos mostram uma tendência à alta flexibilidade abaixo do sítio de início de transcrição. Figura retirada da referência [48].

inicialmente citados em sequências promotoras de humanos depositadas no DBTSS (*Database of Transcription start site*) e constatou que todos eles forneciam o que parece ser uma característica estrutural geral à maioria dos promotores: baixa flexibilidade *upstream* ao TATA-*box* e alta flexibilidade *downstream* ao ponto de início da transcrição (Figura 5).

Fukue *et al.* [20] também utilizou a soma direta das flexibilidades dos *n-mers* para construir perfis de flexibilidade média com os modelos de Brukner *et al.* [59] e de Satchwell *et al.* [60] em sequências promotoras de humanos depositadas EPD. A diferença

foi que ele separou o conjunto total de sequências promotoras em subconjuntos que possuiam apenas TATA-*box* ou apenas INR. Assim como Pedersen *et al.* [48], Fukue *et al.* [20] observou que nestas sequências, a porção *upstream* ao TATA-*box* também tende a ser mais rígida do que a porção *downstream* ao TSS (veja a figura 6, por exemplo), indicando que a flexibilidade do DNA é uma característica estrutural que pode funcionar como um importante marcador em programas de reconhecimento de promotores tal como o *McPromoter* [67] e o *ProStar* [68].

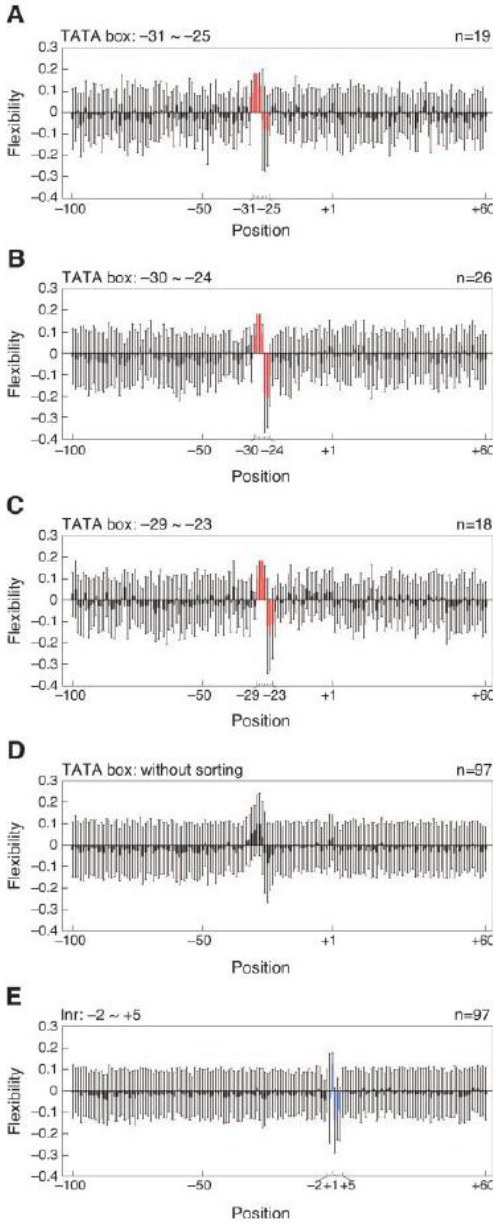


Figura 6

Exemplo de perfis de flexibilidade média das sequências promotoras de humanos do EPD contendo apenas TATA-*box* ou INR, usando o modelo de sensibilidade à DNase I [59]: Valores altos indicam maior flexibilidade. (A), (B) e (C) mostram os resultados para 19, 26 e 18 promotores que possuem apenas TATA-box localizado nas posições -31 a -25, -30 a -24 e -29 a -23, respectivamente. (D) 97 promotores que possuem TATA-*box* localizado entre as posições -35 e -20 foram analisados coletivamente. (E) Resultados de 97 promotores que possuem o INR localizado entre as posições -2 e +5. Figura retirada da referência [20].

4 Motivação e objetivos

Apesar de ser amplamente utilizada, a soma direta das flexibilidades dos *n-mers* nem sempre parece ser a maneira correta de obter a flexibilidade total de uma determinada sequência genômica. O parâmetro utilizado por Packer *et al.* [63], por exemplo, estabelece valores de constante elástica (k) para todas as 136 possíveis combinações dos quatro nucleotídeos do DNA, baseado na energia potencial de superfície de cada combinação. As medidas de raio-X realizadas para cada tetrâmero forneciam informações *estruturais* sobre a molécula de DNA. Neste tipo de parâmetro, a molécula de DNA é vista como um objeto elástico flexível onde cada tetranucleotídeo se compara à uma "mola" individual. Cada uma dessas "molas" é representada por um valor de constante elástica que indica o quão rígida cada mola é e, quanto maior for este valor, maior é a sua rigidez.

Fisicamente, existem duas formas básicas de duas ou mais molas se associarem: em série ou em paralelo, veja a figura 7. Na associação em série, as molas são dispostas linearmente e a constante elástica total ou equivalente *(k_{eq})* da mola resultante é dada pela soma inversa das constantes elásticas de cada mola individual. Já em uma associação em paralelo, as molas apresentam-se empilhadas e o valor de k_{eq} da mola resultante é dada pela soma direta das constantes elásticas de cada mola individual.

Com estas definições em mente, em 2010 nós encontramos um trabalho publicado por Zeng *et al.* [11] que foi o principal motivador da realização do nosso projeto. Neste trabalho, Zeng [11] utilizou a soma direta das constantes elásticas estruturais de Packer [63] para construir um perfil de flexibilidade média das sequências promotoras de humanos, depositadas no DBTSS (Figura 8). De acordo com o que falamos acima, a utilização dos parâmetros de flexibilidade estruturais [63] nos leva a pensar no DNA como sendo uma grande mola composta pela associação de várias molas individuais, representadas pelos tetranucleotídeos. Esses tetranucleotídeos e, consequentemente os *n-mers*, são dispostos

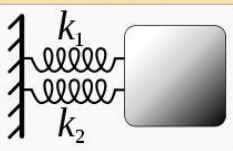
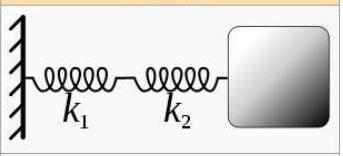
Comparison	In Parallel	In Series
		
Equivalent spring constant	$k_{eq} = k_1 + k_2$	$\frac{1}{k_{eq}} = \frac{1}{k_1} + \frac{1}{k_2}$

Figura 7
As duas formas básicas de duas ou mais molas se associarem: em série e em paralelo. A constante elástica resultante de cada representação, k_{eq}, é obtida por meio das equações mostradas abaixo de cada *box*. Figura modificada, retirada da referência [69].

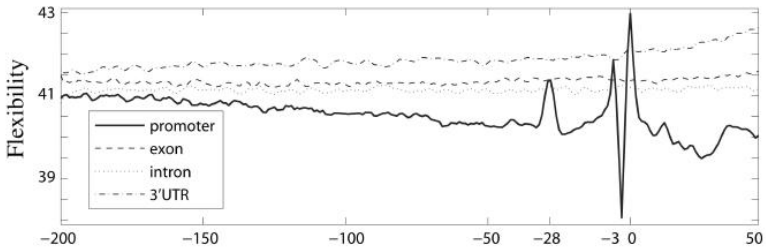


Figura 8
Perfil de flexibilidade média das sequências promotoras de humanos do DBTSS usando o modelo de energia potencial de superfície para tetranucleotídeos [63] O eixo *x* representa as posições dos nucleotídeos nas sequências promotoras (a posição 0 corresponde ao TSS). No eixo *y*, valores mais altos correspondem à sequências supostamente mais rígidas. As posições -28, -3 e 0, compreendidas no núcleo do promotor, possuem propriedades estruturais altalmente distintivas. Uma das regiões mais rígidas é o *6-mer* da posição -28 a -23 e a a região em torno do TSS. Existe uma tendência da porção *upstream* ao TSS ser mais rígida do que a porção *downstream*. Outras regiões genômicas também tiveram seus perfis representados e os valores de flexibilidade encontrados para estas regiões foram muito uniformes. Figura retirada da referência [11].

de forma consecutiva na sequência promotora, o que caracteriza uma associação em série. Logo, o emprego da soma direta de *k* no trabalho de Zeng [11] não foi fisicamente correto.

O valor da flexibilidade de cada posição das sequênciass promotoras foi calculado somando-se as constantes elásticas dos três tetranucleotídeos consecutivos que se sobrepunham a uma sequência de *6-mer*. A equação abaixo resume bem esta metodologia descrita no artigo de Zeng [11],

$$f_i = k_{i,i+3} + k_{i+1,i+4} + k_{i+2,i+5}, \tag{2}$$

onde *i* se refere à posição inicial de cada *6-mer*.

Para exemplicá-la, vamos supor que queremos calcular a flexibilidade total da sequência promotora TATATTA, através da metodologia de Zeng [11]:

- flexibilidade do *6-mer* da primeira posição (TATATT)

$$f_1 = k_{TATA} + k_{ATAT} + k_{TATT}, \tag{3}$$

- flexibilidade do *6-mer* da segunda posição (ATATTA)

$$f_2 = k_{ATAT} + k_{TATT} + k_{ATTA} \tag{4}$$

A flexibilidade total *(k_{eq})* de TATATTA seria dada pela soma direta de f_1 e f_2.

O nosso projeto sugere que, em casos como este, a maneira fisicamente correta de traduzir a flexibilidade total armazenada em uma sequência genômica é através da soma

inversa dos valores de *k* de cada *n-mer*. Sendo assim, os objetivos específicos do nosso projeto podem ser resumidos nos seguintes itens a seguir:

- propor um cálculo de flexibilidade fisicamente correto e desenvolvê-lo em linguagem de programação *Perl*;
- construir perfis de flexibilidade média e mapas de cores para as sequências promotoras reais e randomizadas dos organismos eucariotos considerados com diferentes tamanhos de *n-mers*;
- avaliar o papel da média da flexibilidade na predição de promotores;
- relacionar os valores de flexibilidade média com as sequências de bases nas posições do promotor;
- analisar a frequência dos valores de flexibilidade nas posições do promotor e;
- caracterizar, termodinamicamente, a vizinhança dos genes de microRNAs/mirtrons.

Para atingir estes objetivos, nós dividimos o nosso projeto em 5 sub-projetos:

1. Reprodução da análise de Zeng e introdução da soma inversa de *k* e dos parâmetros de flexibilidade termodinâmicos [70];
2. Aplicação dos parâmetros de flexibilidade termodinâmicos [70] e da soma inversa de *k* nos promotores do EPD;
3. Avaliação da média da flexibilidade como preditor de promotores;
4. Aplicação dos parametros de flexibilidade termodinâmicos [70] e da soma inversa de *k* nas sequências *"core-less"* de humanos e;
5. Caracterização termodinâmica da vizinhança dos microRNAs/mirtrons de *Drosophila melanogaster* e *Caenorhabditis elegans*.

5 Metodologia

5.1 Cálculo fisicamente correto da constante elástica equivalente, k_{eq}

Na seção anterior, foi mostrado que os *n-mers* de uma sequência promotora estão dispostos de forma consecutiva, o que caracteriza uma associação em série. Dessa forma, ao utilizar parâmetros de constante elástica para predizer a flexibilidade total de uma sequência genômica, o ideal é utilizar a soma inversa dos valores de k de cada *n-mer*. Portanto, a equação fisicamente correta para obter a constante elástica equivalente de cada posição da sequência promotora, k_{eq}, é

$$\frac{1}{k_{1,N}} = \sum_{i=1,\ldots,N-1} \frac{1}{k_{i,i+1}} \tag{5}$$

.

Para exemplificá-la, vamos considerar a mesma sequência TATATTA para mostrar como seria o cálculo da flexibilidade de cada posição por meio da soma inversa:

- flexibilidade do 6-*mer* da primeira posição (TATATT)

$$1/f_1 = 1/k_{TATA} + 1/k_{ATAT} + 1/k_{TATT}, \tag{6}$$

- flexibilidade do 6-*mer* da segunda posição (ATATTA)

$$1/f_2 = 1/k_{ATAT} + 1/k_{TATT} + 1/k_{ATTA}. \tag{7}$$

Logo, a flexibilidade total *(k_{eq})* da sequência TATATTA seria dada pela soma inversa de $1/f_1$ e $1/f_2$.

5.2 Dados utilizados

5.2.1 Parâmetros estruturais e termodinâmicos

Os valores de constante elástica estruturais (k) dos 136 tetranucleotídeos foram obtidos do artigo de Packer [63], veja a tabela 2. Diferentemente de Packer [63], Weber [70] utilizou um modelo físico-estatístico para calcular as constantes elásticas de dinucleotídeos baseado em valores de temperatura de denaturação da molécula de DNA. Como estes parâmetros refletem a termodinâmica da molécula, vamos nos referir à estes como parâmetros *termodinâmicos*. Devido à simetria que existe entre alguns pares de bases, o cálculo foi realizado sobre 10 pares canônicos: AA(=TT); AT; AC(=GT);

AG(=CT); TA; GA(=TC); CA(=TG); CG; GC; GG(=CC) e os resultados obtidos estão mostrados na tabela 3.

5.2.2 Sequências promotoras

Para reproduzir a análise de Zeng [11], todas as 30046 sequências promotoras de humanos foram extraídas do banco de dados DBTSS. O DBTSS, banco de dados utilizado por Zeng [11], contêm sequências promotoras detectadas por meio do mapeamento de clones de cDNAs de humanos e ratos [9, 72]. Devido à este método, a quantidade e a variedade dos sítios de início de transcrição depositados no DBTSS é imensa. Para realizar as outras análises, nós escolhemos o EPD, um banco de dados de promotores eucarióticos reconhecidos apenas pela pol-II. No EPD, os sítios de início de transcrição são rigorosamente selecionados e, por conta disso, as informações contidas neste banco de dados não são redundantes [73]. Levando em consideração a importância biológica e a quantidade de promotores, três organismos tiveram suas sequências promotoras coletadas do EPD: *Homo sapiens*, *Drosophila melanogaster* e *Oriza sativa*. A quantidade de promotores de cada um deles está informado na tabela 4.

O arquivo dos promotores de humanos coletados do EPD foi subdividido em vários outros arquivos contendo apenas sequências com TATA-*box*, INR, DPE e BRE. O restante das sequências representavam os promotores *"core-less"* de *H. sapiens* e foi sobre estes arquivos que aplicamos os parâmetros de flexibilidade termodinâmicos [71] e a soma inversa da flexibilidade. Todas as sequências promotoras extraídas continham 150 pb *upstream* e 50 pb *downstream* ao TSS, totalizando 201 pb. Uma janela deslizante de tamanhos variados deslizava de posição em posição para calcular o k_{eq} de cada *n-mer* da sequência promotora.

5.3 Randomização das sequências promotoras

Para avaliar se a flexibilidade média na região promotora do DNA depende ou não da sua disposição de bases, nós construimos um arquivo para cada um dos três organismos selecionados (coluna 3 da tabela 4). Na randomização, as bases de cada sequência são reordenadas nas 201 posições, destruindo qualquer informação de significado biológico mas mantendo intacta outras informações como o conteúdo relativo de A, C, G e T. O procedimento é repetido 8 vezes para cada sequência promotora e incluído no arquivo, ou seja, cada promotor aparece com 8 cópias aleatorizadas.

Tetrâmero	k	Tetrâmero	k	Tetrâmero	k	Tetrâmero	k
TACA	1,9	CCAG	9,1	GAGC	11,9	CAGC	17,1
CATA	3,2	CGAC	9,2	CCGA	12,1	GACG	17,3
TATA	3,6	AACA	9,2	CTAC	12,1	AACC	17,3
TGCA	3,8	ACAA	9,3	GGCA	12,1	CAAC	17,5
GCGC	3,9	ATAC	9,4	AGCT	12,2	AGAA	17,7
CACG	4,1	AAGT	9,6	TGAA	12,5	CTAG	17,8
GAGT	4,4	CGCA	9,6	CGAG	12,6	AGAG	17,9
CACT	4,8	AGCG	9,7	AATG	12,7	GGAC	18,0
CTAA	5,0	AGCA	9,9	TAGA	13,1	TAGT	18,1
ATAA	6,0	TCAC	9,9	CCAA	13,1	CGAA	18,1
TACG	6,2	CGCG	10,0	GGAG	13,3	GGGT	18,3
CACA	6,6	TCGA	10,2	ACAT	13,5	TAAA	18,8
GCAC	6,6	TGGC	10,3	CAGA	13,5	GGCC	19,1
GCGA	7,2	GATA	10,4	GAGG	13,6	CGAT	19,2
GCAG	7,2	AAGG	10,6	TGAC	13,7	GAAA	19,3
ACGG	7,3	TGGT	10,6	GGGA	13,8	TTAA	19,4
CATG	7,7	GGAA	10,6	CATC	13,9	GGGC	19,6
CCAC	7,9	GTAA	10,6	AAGC	14,3	TAAG	19,8
CGGA	7,9	TCAT	10,7	CGGC	14,5	AGAC	20,1
TAGG	7,9	AGGG	10,8	TGAG	14,5	CAAG	20,2
GTAC	8,0	CAGG	10,8	AGAT	14,5	TAAC	20,7
GCAT	8,0	TGGG	11,1	TAAT	14,7	GAAT	21,3
ACGA	8,0	ACAG	11,2	AATA	14,8	GAAC	21,5
GACA	8,4	ATAG	11,4	AACG	14,9	AAAT	21,7
CGCC	8,6	CCGC	11,4	CGGT	15,5	CAAT	22,3
TACT	8,6	GGAT	11,4	AACT	15,5	GAGA	22,7
TGGA	8,6	ACGT	11,5	AGGT	15,7	AAAA	23,8
ATAT	8,6	TAGC	11,6	AGGC	15,9	CAAA	24,3
GCAA	8,7	TCAA	11,6	AGCC	16,0	AAAG	24,5
CACC	8,7	CCAT	11,7	ACAC	16,2	GAAG	24,9
TACC	8,8	AGGA	11,7	GGGG	16,4	AAGA	25,3
TGAG	8,9	ACGC	11,7	CCGG	16,6	AATC	26,8
TGAT	9,0	CAGT	11,9	GACT	16,9	AATT	26,9
GACC	9,1	CGGG	11,9	GATC	17,0	AAAC	27,2

Tabela 2
Parâmetros estruturais de Packer em $\mathrm{kJ \cdot mol^{-1} \cdot Å^{-2}}$: Valores mais altos de k representam tetrâmeros mais rígidos [63].

Dímero	k	Dímero	k
AA	0,0248659	AC	0,0230182
AG	0,0220903	AT	0,0109925
CA	0,0315708	CC	0,0209859
CG	0,0245137	GA	0,0312279
GC	0,0359905	TA	0,0353204

Tabela 3
Parâmetros de flexibilidade termodinâmicos para pareamentos canônicos de bases, dados em $\mathrm{eV \cdot Å^{-2}}$: Valores mais altos de k representam dímeros mais rígidos [71].

Organismo	nº de promotores EPD	nº de promotores considerados	quantidade de randomizados
H. sapiens	1871	1861	14888
D. melanogaster	1922	1922	15376
O. sativa	13046	13038	104304
"core-less" de *H. sapiens*	-	1539	12312

Tabela 4
A coluna 2 representa a quantidade de promotores extraídos do EPD, para cada organismo listado na coluna 1 da tabela. Devido a presença de bases ainda não identificadas nas sequências extraídas, nem todos promotores puderam ser utilizados em nossas análises.

5.4 Obtenção da flexibilidade média, desvio padrão e mapas de cores

Conforme falamos, todas as sequências promotoras extraídas continham 150 pb *upstream* e 50 pb *downstream* ao TSS, totalizando 201 pb. Cada posição teve seu valor de flexibilidade média calculado por meio de janelas deslizantes de diferentes tamanhos que deslizavam de posição em posição. Todos os valores de k_{eq} de uma mesma posição eram somados e a soma resultante era dividida pela quantidade de promotores de cada organismo em questão. O valor obtido por meio desta divisão representava a flexibilidade média de cada posição da sequência promotora. O desvio padrão médio de cada posição foi calculado para mostrar o quanto os valores de flexibilidade se aproximavam ou se afastavam do valor de flexibilidade média obtido para cada posição.

Para entender melhor o comportamento da flexibilidade nestas sequências, nós armazenamos o valor mínimo e máximo de flexibilidade média encontrado para cada arquivo e, a partir do valor mínino, fomos contabilizando a frequência de cada um destes valores para cada posição da sequência promotora até chegar no valor máximo de flexibilidade média encontrado. Os resultados desta análise pode ser visto através dos mapas de cores que foram construídos para fornecer todos os valores possíveis de flexibilidade de cada arquivo e a taxa de ocorrência de cada um deles em todas as posições.

5.5 Modelo de estiramento

A visualização dos resultados encontrados em cada sub-projeto foi feita através de gráficos de perfis de flexibilidade média. No eixo x dos gráficos foram representadas as posições das sequências promotoras, que variam de -150 a +50 e, no eixo y foram representados os valores de flexibilidade média, dados em S. S se refere à conversão da flexibilidade em estiramento, realizada através da equação

$$S_{eq} = k_{eq}(Nx_0). \tag{8}$$

$N = n - 1$ e $x_0 = 0{,}33$ Å se referem, respectivamente, ao tamanho dos *n-mers* e à distância média de dois pares de bases vizinhos. Esta conversão foi necessária para poder comparar a flexibilidade média dos diferentes tamanhos de *n-mers*. Quanto maior o valor de S, maior é a rigidez da sequência.

6 Resultados e discussão

6.1 Reprodução da análise de Zeng e introdução da soma inversa de k e dos parâmetros de flexibilidade termodinâmicos

A utilização do mesmo banco de dados e do mesmo parâmetro de flexibilidade foram fundamentais para que conseguíssemos reproduzir, com fidelidade, toda a análise da soma direta de k realizada no trabalho de Zeng [11], veja o gráfico 9a. As 30046 sequências promotoras de humanos foram extraídas do banco de dados DBTSS e os valores de constante elástica (k) de cada tetranucleotídeo foram provenientes do artigo do Packer do ano 2000 e mostrados na tabela 2. Para calcular valor da flexibilidade de cada posição, k_{eq}, uma janela de tamanho igual a 6 deslizava de posição em posição e os três tetranucleotídeos consecutivos que se sobrepunham à cada janela de *6-mer* tiveram seus valores de constantes elásticas somados de forma direta.

No tipo de representação utilizada por Zeng [11], cada tetranucleotídeo da sequência de *6-mer* apresenta três nucleotídeos que são idênticos ao tetranucleotídeo anterior. Para minimizar o efeito redundante da sobreposição de nucleotídeos, nós sugerimos que o cálculo da flexibilidade de cada posição de uma sequência promotora deveria ter sido feito considerando uma sequência de *7-mer*, conforme mostra a equação

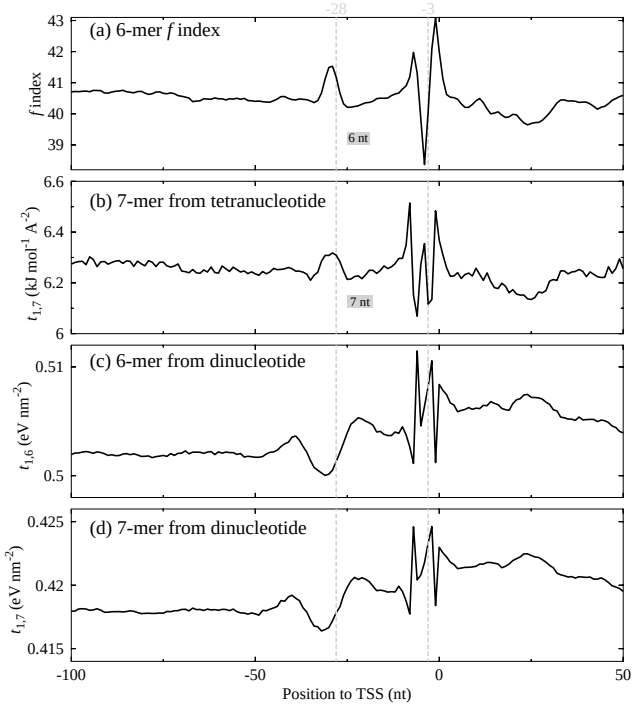


Figura 9
Perfis de flexibilidade média usando janela (a) *6-mer* para os parâmetros de flexibilidade estruturais [63] assim como Zeng [11] utilizou em seu trabalho, (b) *7-mer* para os parâmetros de flexibilidade estruturais [63], (c) *6-mer* para os parâmetros de flexibilidade termodinâmicos [71] e (d) *7-mer* para os parâmetros de flexibilidade termodinâmicos [71]. Figura retirada do nosso artigo [74], reproduzido na seção C do apêndice.

$$1/f_i = 1/k_{i,i+3} + 1/k_{i+3,i+6}, \tag{9}$$

onde i se refere à posição inicial de cada *7-mer*. Nesta representação, cada *7-mer* é composto por dois tetranucleotídeos que se sobrepõem apenas pelo nucleotídeo localizado na quarta posição. Veja abaixo a diferença:

- composição direta do ATTTCG, conforme Zeng utilizou: ATTT, TTTC e TTCG. Observe que o segundo tetranucleotídeo se mantêm ligado ao primeiro através dos três nucleotídeos finais presentes no primeiro tetranucleotídeo. A mesma correlação pode ser feita entre o terceiro e o segundo tetranucleotídeo.

- composição física do ATTTCGT, conforme sugerimos: ATTT e TCGT. Neste caso, o segundo tetranucleotídeo se mantêm ligado ao primeiro apenas pelo último nucleotídeo presente no primeiro tetranucleotídeo.

A principal diferença gerada por esta modificação foi a existência de uma depressão extra próxima ao sítio de início de transcrição, indicando que, ao invés de uma, as sequências promotoras possuem duas regiões mais flexíveis próximas ao TSS, como podemos ver no gráfico 9b.

Outra modificação que também sugerimos foi a realização de uma análise utilizando os parâmetros de flexibilidade termodinâmicos mostrados tabela 3 [71] e a soma inversa de k (seção 5.1) no lugar dos parâmetros de flexibilidade estruturais [63] e da soma direta, respectivamente. O gráfico 9c e 9d retratam exatamente o comportamento diferenciado do perfil de flexibilidade média após estas duas modificações, com *n-mer* de tamanho 6 e 7. A utilização deste tipo de paramêtro considera que o DNA se comporta como uma grande mola composta por várias molas individuais dispostas em série. Fisicamente, o valor da constante elástica resultante k_{eq} de duas ou mais molas associadas em série é dado pela soma inversa dos valores de k de cada mola. Logo, a soma direta dos valores de k de cada mola não fornece dados representativos da física do DNA. Independente do tamanho de *n-mer* considerado, nós podemos ver que o uso da soma inversa de k nas regiões promotoras geraram conclusões contrárias ao que é conhecido sobre as características estruturais gerais das sequências promotoras: em ambos os gráficos, a porção *downstream* ao TSS apresenta-se mais rígida do que a porção *usptream* ao TATA-*box* e a posição em torno de -28 é uma das regiões mais flexíveis.

6.2 Aplicação dos parâmetros de flexibilidade termodinâmicos e da soma inversa de k nos promotores do EPD

Os resultados encontrados no subprojeto 1 nos fez aplicar os parâmetros de flexibilidade termodinâmicos e o modelo de soma inversa de *k* em sequências promotoras de três organismos eucariotos presentes no banco de dados EPD: *H. sapiens*, *D. melanogaster* e *O. sativa*, seção 5.2. Os valores de *k* da tabela 3 foram úteis para predizer a flexibilidade média das regiões promotoras e a sua aplicação trouxe a possibilidade de calcular a flexibilidade média utilizando vários tamanhos de *n-mer*. Os arquivos com as sequências randomizadas de cada organismo (seção 5.3) também foram submetidos à mesma análise e, devido à quantidade de dados gerados, vamos exemplificar os resultados apenas para *n-mer* de tamanho 6, 10, 14 e 20. Cada gráfico de perfil de flexibilidade média é acompanhado por um gráfico que representa o desvio padrão médio da flexibilidade de cada posição.

Nos gráficos 10a, 11a e 12a temos os perfis de flexibilidade média calculado para as sequências promotoras reais de *H. sapiens*, *D. melanogaster* e *O. sativa*, respectivamente. É possível perceber que a variação do tamanho dos *n-mers* não gerou mudanças bruscas no perfil de flexibilidade média encontrado para cada organismo e, independente do seu tamanho, o comportamento da flexibilidade foi muito semelhante em todos os organismos: baixa flexibilidade *downstream* e maior flexibilidade *upstream* ao TATA. Algumas características do subprojeto 2 foram muito parecidas com as que encontramos no subprojeto 1: em todos os organismos, a região em torno da posição -28 e do TSS também apresentaram-se bem flexível quando comparado com o restante das posições (ver figura 9c e d). A barra amarela presente nos gráficos representa a localização do elemento TATA-*box* que ocorre em torno da posição -28 e, como podemos ver, este elemento apresenta-se bastante flexível.

Por outro lado, a realização do mesmo tipo de análise com as sequências promotoras randomizadas revelou que, independente do tamanho do *n-mer* e do organismo em questão, o comportamento da flexibilidade média apresentou-se bastante uniforme, como mostram os gráficos 10c, 11c e 12c. Apesar de ter sido mantida a frequência das bases, a randomização destruiu toda e qualquer informação biológica contida em cada sequência promotora e os resultados encontrados reafirmam que a predição da flexibilidade depende diretamente da organização dessas bases.

Os gráficos 10b, 11b e 12b trazem os perfis de desvio padrão médio da flexibilidade das sequências promotoras reais de cada organismo estudado e, de acordo com os gráficos, todos eles são bastante parecidos com os perfis de desvio padrão médio da flexibilidade encontrados para as sequências randomizadas, mostrados as figuras 10d, 11d e 12d.

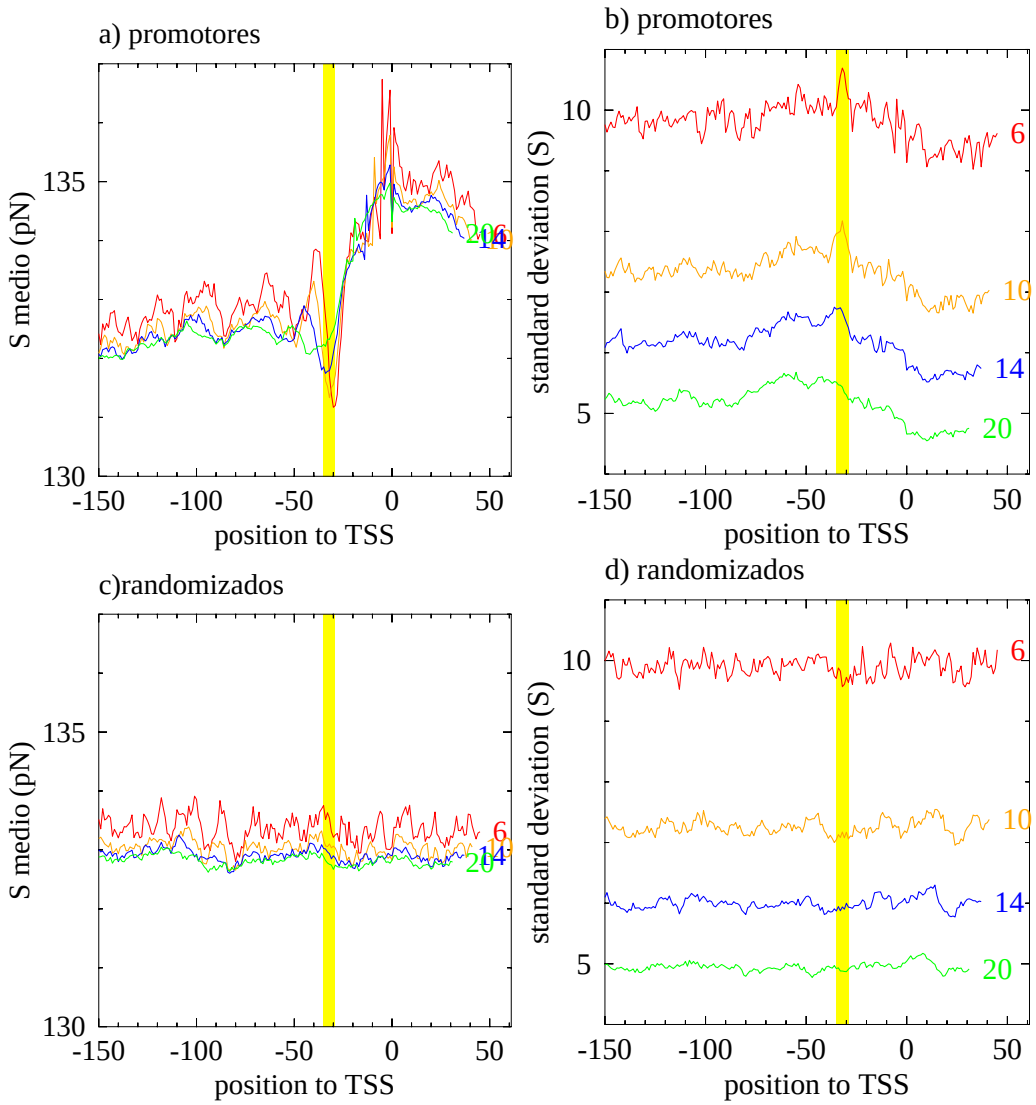


Figura 10

(a) Perfil de flexibilidade média das sequências promotoras reais de *H. sapiens* e, em (b), o desvio padrão médio da flexibilidade das sequências promotoras reais de *H. sapiens*. (c) Perfil de flexibilidade média das sequências promotoras randomizadas de *H. sapiens* e em (d), o desvio padrão médio da flexibilidade das sequências randomizadas *H. sapiens*. O cálculo da flexibilidade foi realizado com vários tamanhos de *n-mer*. Em ambos os gráficos, cada cor representa um tamanho de *n-mer*: vermelho (n=6), laranja (n=10), azul (n=14) e verde (n=20).

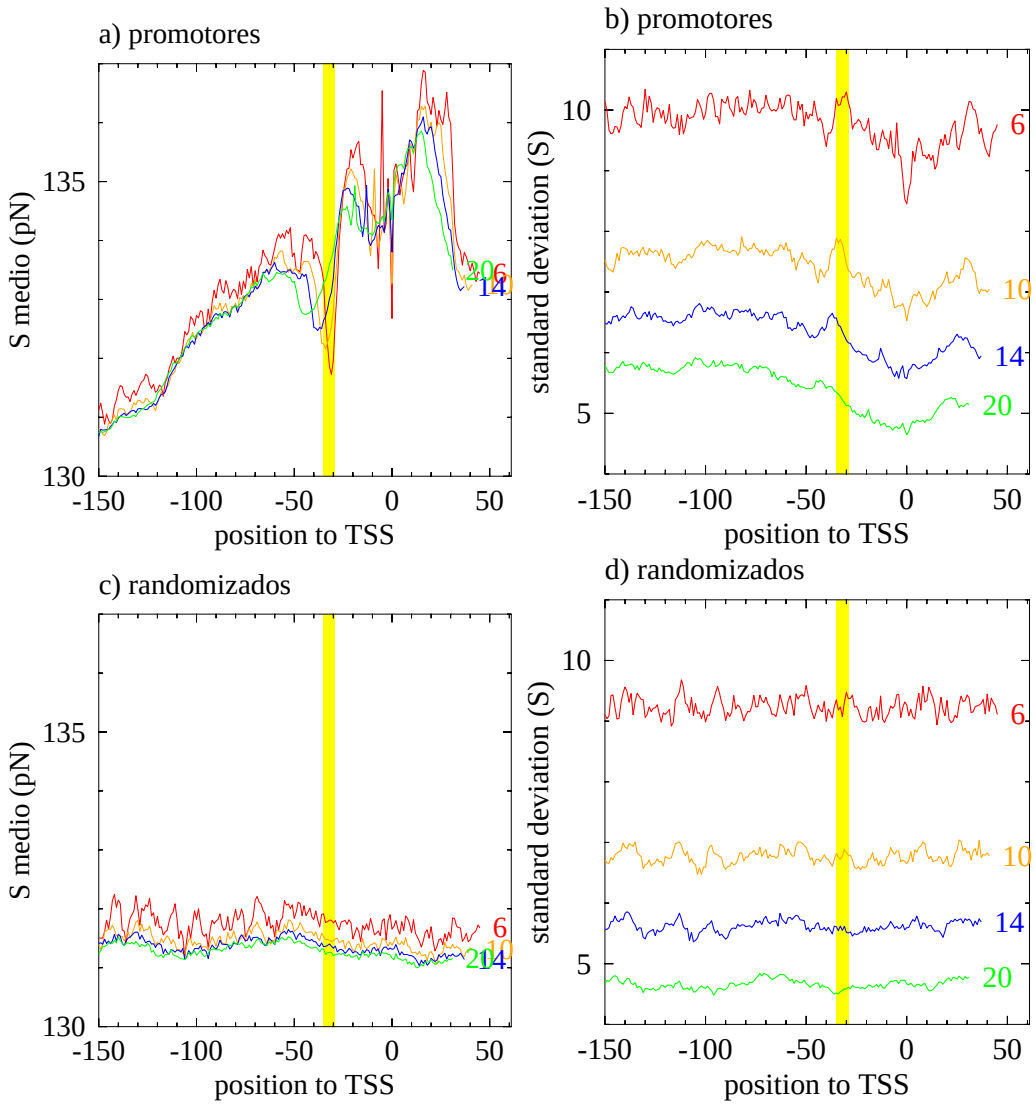


Figura 11

(a) Perfil de flexibilidade média das sequências promotoras reais de *D. melanogaster* e, em (b), o desvio padrão médio da flexibilidade das sequências promotoras reais de *D. melanomo mostraogaster*. (c) Perfil de flexibilidade média das sequências promotoras randomizadas de *D. melanogaster* e em (d), o desvio padrão médio da flexibilidade das sequências randomizadas *D. melanogaster*. O cálculo da flexibilidade foi realizado com vários tamanhos de *n-mer*. Em ambos os gráficos, cada cor representa um tamanho de *n-mer*: vermelho (n=6), laranja (n=10), azul (n=14) e verde (n=20).

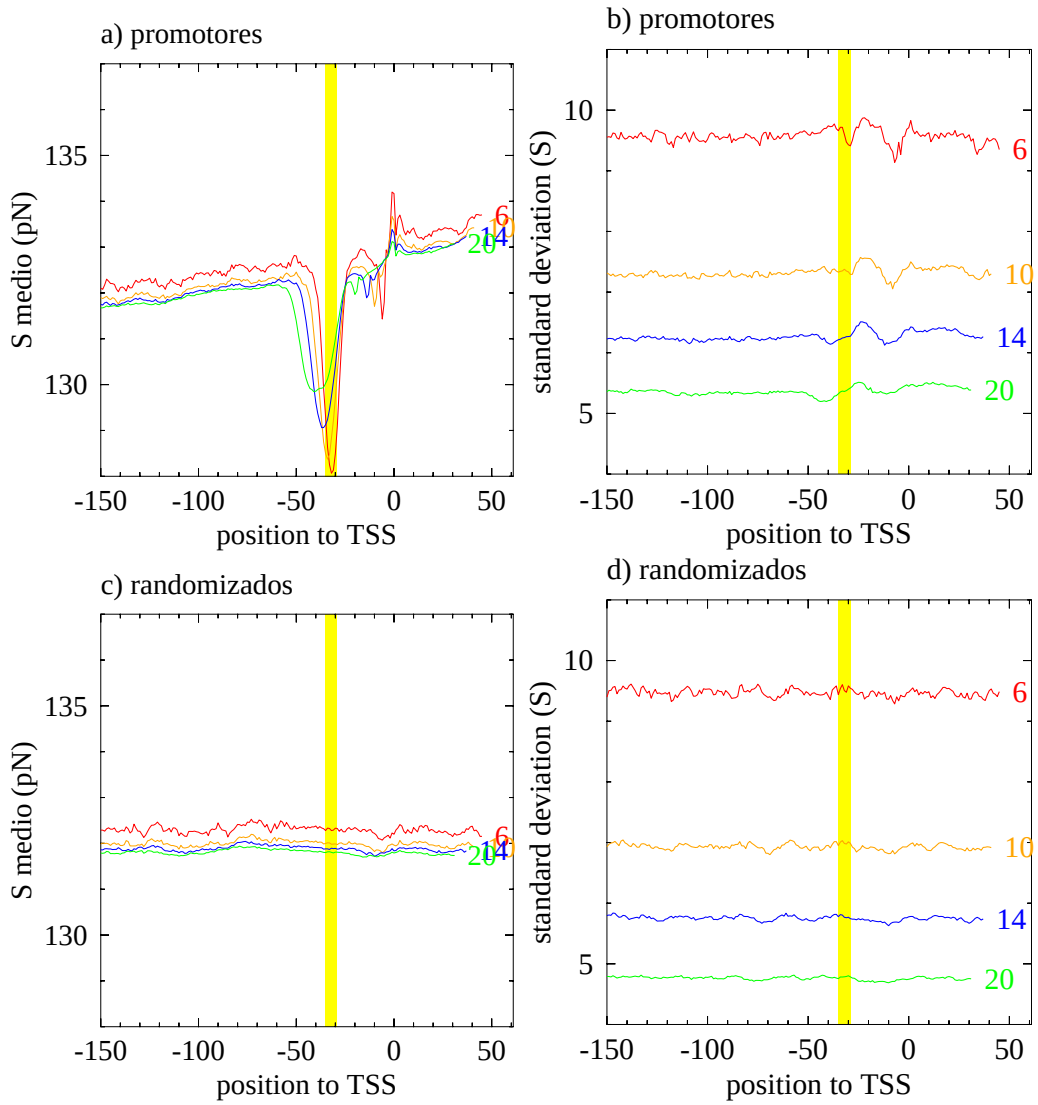


Figura 12

(a) Perfil de flexibilidade média das sequências promotoras reais de *O. sativa* e, em (b), o desvio padrão médio da flexibilidade das sequências promotoras reais de *O. sativa*. (c) Perfil de flexibilidade média das sequências promotoras randomizadas de *O. sativa* e em (d), o desvio padrão médio da flexibilidade das sequências randomizadas *O. sativa*. O cálculo da flexibilidade foi realizado com vários tamanhos de *n-mer*. Em ambos os gráficos, cada cor representa um tamanho de *n*: vermelho (*n*=6), laranja (*n*=10), azul (*n=14) e verde (n=20).*

6.3 Avaliação da média da flexibilidade como preditor de promotores

Apesar de ser a medida de posição mais utilizada, a média de qualquer conjuntos de valores nem sempre é suficiente para descrever um grupo de dados pois a amplitude de variação de seus dados pode ser muito alta. No caso do nosso projeto, cada valor do eixo y dos gráficos de perfis de flexibilidade média representa a flexibilidade de 1000 ou mais posições, o que não significa que todas as sequências vão apresentar este mesmo valor em suas posições. A figura 13a, por exemplo, mostra que o perfil de flexibilidade de uma sequência promotora qualquer de *H. sapiens*, representado pela linha rosa, não apresenta um perfil que se assemelha ao perfil de flexibilidade média calculado para todas as sequências promotoras de humanos. De fato, o perfil deste promotor não está sequer contido nos limites estabelecidos pelo desvio padrão (faixa cinza da figura 13a). Quando randomizamos este mesmo promotor (figura 13b), vemos que o seu perfil de flexibilidade também se distancia bastante do perfil de flexibilidade média de todas as sequências promotoras de *H. sapiens*. Em algumas posições, nós esperávamos que o desvio padrão médio da flexibilidade das sequências promotoras reais fosse menor do que o esperado para as mesmas posições das sequências randomizadas. Entretanto, os valores encontrados para estes dois arquivos foram muito parecidos, o que inviabiliza a utilização da flexibilidade média na busca por promotores. Ou seja, apenas pela flexibilidade e o desvio padrão calculados a partir de um conjunto de promotores, não somos capazes de distinguir um promotor pertencente ao próprio conjunto de uma sequência randomizada qualquer.

Diante da inadequação da média como preditor, nós resolvemos estimar a taxa de ocorrência de todos os valores de flexibilidade encontrados para cada tamanho de *n-mer* que foi utilizado no cálculo de flexibilidade média das sequências promotoras reais e randomizadas dos organismos selecionados do EPD. Para visualizar os resultados, foram construídos mapas de cores, onde a frequência dos valores de flexibilidade em cada posição foi relacionada com as cores presentes na barra lateral dos mapas. Quanto mais próximo azul, mais frequente é aquele valor em determinada posição e, quanto mais próximo do vermelho, menor é a sua frequência. Por exemplo, no mapa 14 (a), cerca de 10% das sequências promotoras de humanos apresentam o valor de 120 pN em torno da posição -25. No corpo deste texto, nos vamos apresentar os mapas de cores das sequências promotoras reais e randomizadas de *H. sapiens*, *D. melanogaster* e *O. sativa* com *n-mer* iguais à 6 e 20 e, na seção A do apêndice estão os mapas de cores adicionais destes mesmos organismos com *n-mer* iguais a 10 e 14.

Em essência, os mapas de cores que são discutidos nesta seção representam a probabilidade de encontrar um dado valor de flexibilidade numa certa posição da sequência

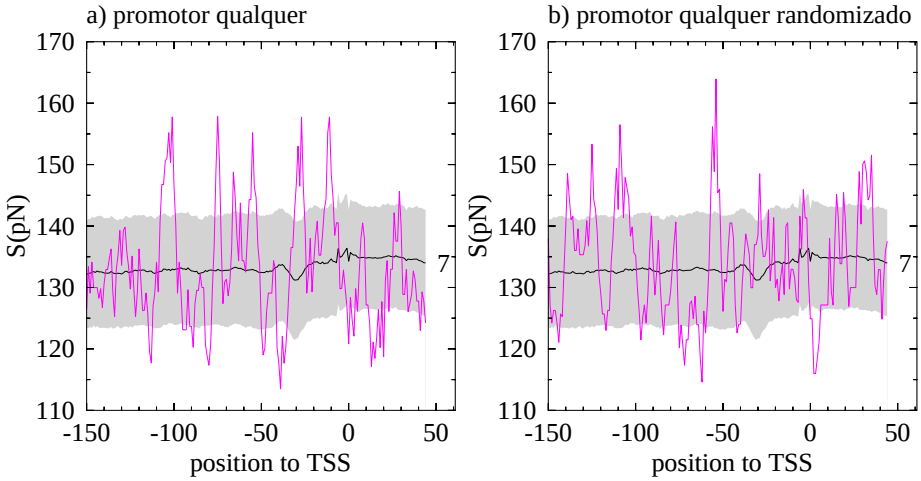


Figura 13
(a) Perfil de flexibilidade de um promotor qualquer de *H. sapiens* (linha rosa) comparado ao perfil de flexibilidade média de todos os promotores de *H. sapiens* com *n-mer* igual à 7 (linha preta). A faixa em cinza representa os limites do desvio padrão médio da flexibilidade calculado com tamanho de *n-mer*. (b) A mesma comparação feita em (a), porém, com o promotor qualquer randomizado.

promotora. Ou seja, os mapas são a visualização desta probabilidade. Diferenças entre estes mapas para os correspondentes de sequências randomizadas indicam se será possível usar estas probabilidades como preditores de sequências promotoras. Apesar deste tipo de probabilidade ser usado em diversos trabalhos de predição de promotores (por exemplo [11]), não há uma análise sistemática em termos de tamanho de *n-mer* e nem em relação à sequências randomizadas.

Tanto nos mapas de cores dos promotores reais quanto nos dos randomizados, a primeira característica que nos chama a atenção são as faixas coloridas que aparecem na horizontal dos mapas, figuras 14, 15 e 16. Estas faixas ficam mais definidas quando aumentamos o tamanho do *n-mer* e a presença delas revelam que as posições das sequências promotoras analisadas se comportam de maneira semelhante quanto à flexibilidade, ou seja, os valores mais altos e mais baixos aparecem com uma frequência reduzida ao passo que os valores intermediários aparecem com uma frequência maior.

Outra característica que observamos foi que, nos mapas de cores das sequências randomizadas dos três organismos selecionados, a frequência dos valores de flexibilidade foram bastante uniformes ao longo das posições, veja os mapas (b) e (d) das figuras 14, 15 e 16. Nos mapas de cores dos promotores reais, entretanto, nós podemos ver que a frequência dos valores de flexibilidade sofrem uma certa variação abaixo da posição -50, em geral (mapas (a) e (c) das figuras 14, 15 e 16). Em humanos, repare que abaixo da posição -30, a frequência dos valores mais baixos de flexibilidade parecem com uma

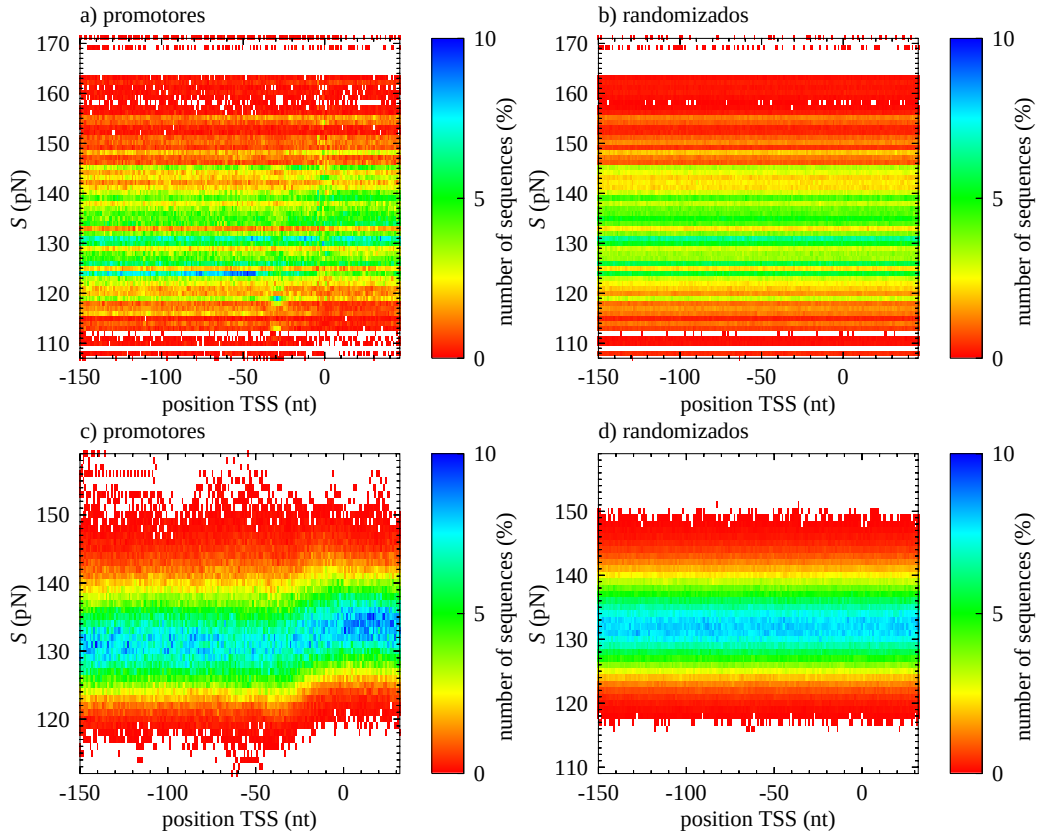


Figura 14
Mapas de cores das sequências promotoras de *H. sapiens* com *n-mer* de tamanho (a) 6 e (c) 20. (b) e (d) representam os mapas das sequências randomizadas *n-mer* de tamanho 6 e 20, respectivamente.

frequência ainda mais reduzida quando comparado com as posições acima da posição -30. No caso da *D. melanogaster*, esta variação ocorre gradualmente mas também é nítido que abaixo da posição -30 os valores de flexibilidade mais baixos passam a ser menos frequentes do que acima desta posição. O mesmo é observado nos mapas de cores de *O. sativa*, figura 16.

Exceto com $n = 6$, a utilização de vários tamanhos de *n-mer* nesta análise não foi capaz de gerar mudanças bruscas na frequência dos valores de flexibilidade encontrado para as sequências randomizadas dos organismos analisados, mapas (b) e (d) das figuras 14, 15 e 16. Em contrapartida, quando variamos o tamanho do *n-mer* das sequências promotoras reais, vemos que existem alguns valores de flexibilidade que apresentam frequência diferenciada em determinadas posições, repare nos mapas (a) e (c) das figuras 14, 15 e 16. Com *n-mer* igual a 6, por exemplo, nós podemos observar que o valor de flexibilidade próximo à 124 *S* é bastante frequente em torno da posição -50 das sequências promotoras reais de *H. sapiens*. Já com *n-mer* igual a 20, os valores de flexibilidade próximos à 130

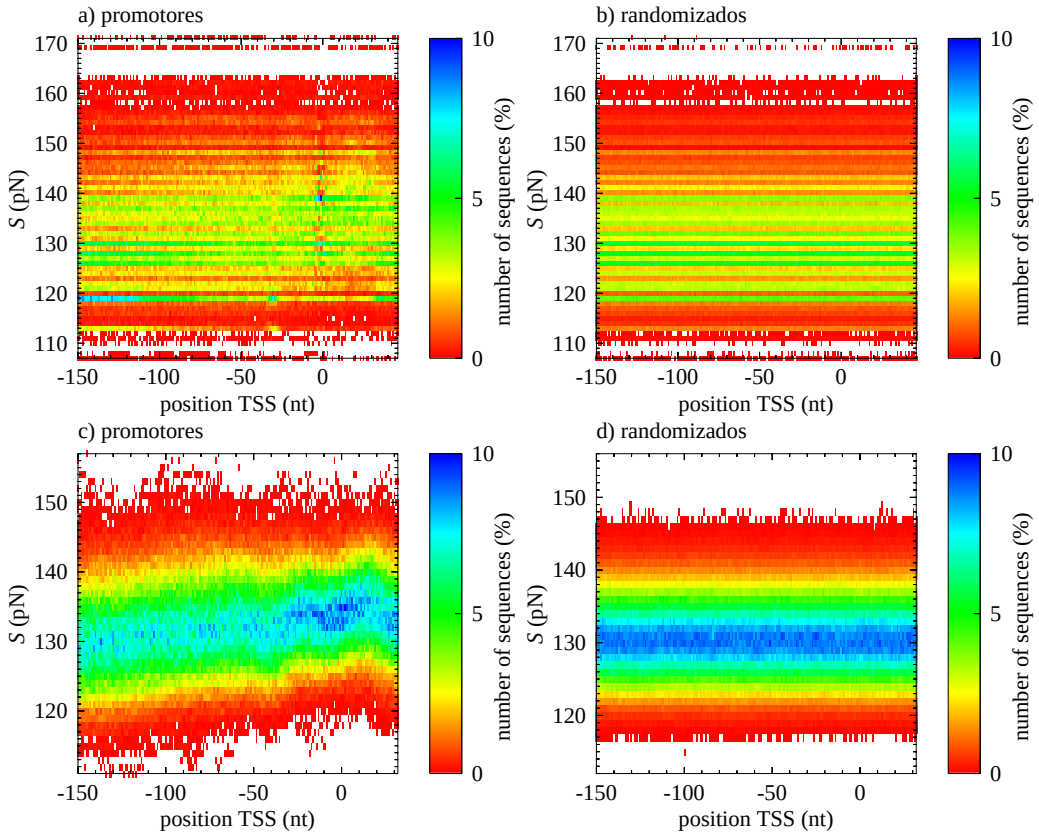


Figura 15
Mapas de cores das sequências promotoras de *D. melanogaster* com *n-mer* de tamanho (a) 6 e (c) 20. (b) e (d) representam os mapas das sequências randomizadas *n-mer* de tamanho 6 e 20, respectivamente.

S aparecem frequentes em várias posições das sequências promotoras reais de *H. sapiens*, porém, é abaixo do TSS que a frequência destes valores se destaca.

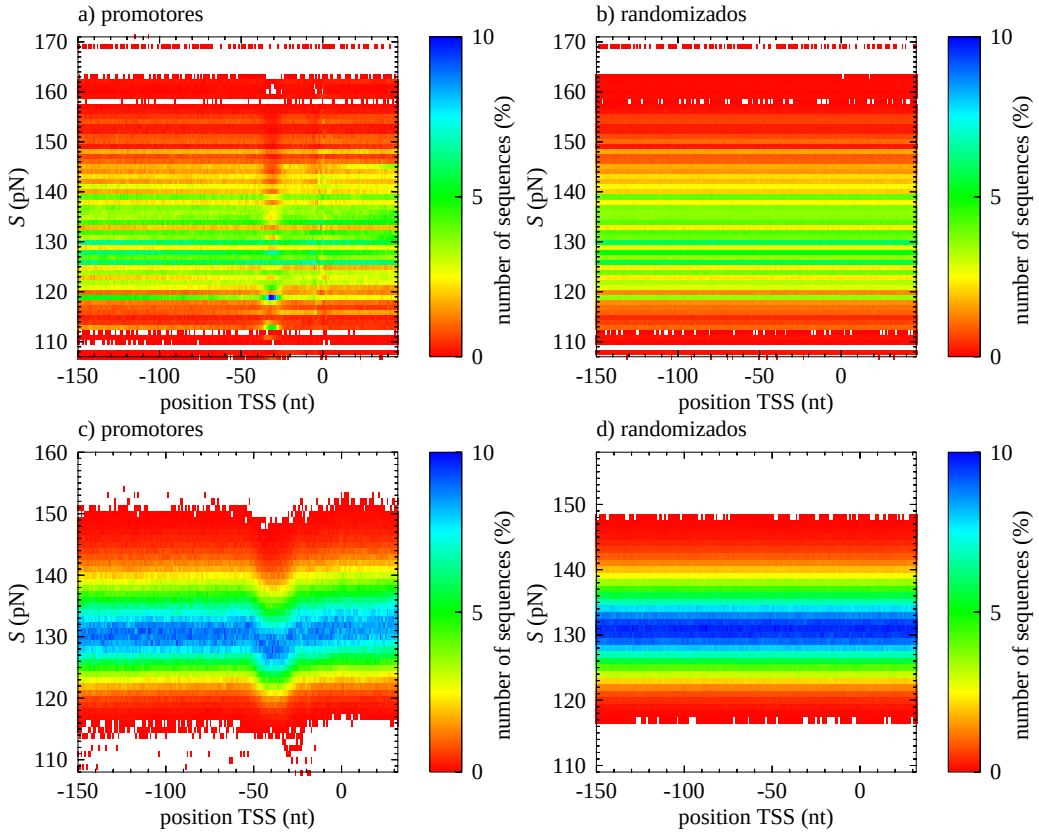


Figura 16
Mapas de cores das sequências promotoras de *O. sativa* com *n-mer* de tamanho (a) 6 e (c) 20. (b) e (d) representam os mapas das sequências randomizadas *n-mer* de tamanho 6 e 20, respectivamente.

6.4 Aplicação dos parâmetros de flexibilidade termodinâmicos e da soma inversa de k nas sequências *"core-less"* de humanos

Esta parte do nosso projeto se baseou em analisar o comportamento da flexibilidade média nos promotores *"core-less"* de humanos através da soma inversa das constantes elásticas de cada *n-mer* (seções 5.2 e 5.1). Conforme Fukue *et al.* [38] estabeleceu em um de seus trabalhos, este grupo de promotores não possuem qualquer tipo de elemento regulatório e, mesmo assim, são reconhecidos pelo aparato enzimático da transcrição, levando a crer que as propriedades físicas nos promotores *"core-less"* são realmente relevantes. Apesar da ausência de qualquer elemento regulatório conhecido, o comportamento do perfil de flexibilidade média que encontramos para este grupo de promotores foi muito parecido com o encontrado para os outros conjuntos de promotores, veja o gráfico à esquerda figura 17. Na região em torno da posição -28, conhecida como a região que possui o elemento TATA-*box*, a flexibilidade começa a aumentar após uma depressão acentuada em torno da posição -25. O pico encontrado próximo ao TSS, conhecida como a região que engloba o INR, revela que a flexibilidade média em torno desta posição também é alta. No gráfico à direita da figura 17, um dos resultados que se destacou foi o reduzido valor de desvio padrão médio da flexibilidade com *n-mer* igual a 14 próximo à posição do TATA-*box*.

Também estimamos a frequência de todos os valores de flexibilidade encontrados para cada *n-mer* utilizado no cálculo de flexibilidade média dos promotores *"core-less"*. Os mapas de cores obtidos revelam que, com *n-mer* igual à 6, valores de flexibilidade próximos de 125 *S* são abundantes em torno da posição -50, assim como foi visto no mapa de cores dos promotores de humanos com o mesmo tamanho de *n-mer* (figura 14). Já com *n-mer* igual a 20, é possível perceber uma depressão mais significativa da frequência de flexbilidade próximo da poisção -50. Os resultados encontrados para os promotores *"core-less"* nos ajudam a entender o porquê que estes promotores não deixam de ser reconhecidos pelo aparato enzimático da transcrição, mesmo sem a presença dos elementos regulatórios conhecidos. Provavelmente, são as características físicas dos promotores *"core-less"*, tal como a flexibilidade do DNA, que sinalizam para os fatores de transcrição reconhecê-los e dar início ao processo transcricional. Além disso, nós não descartamos a hipótese de que possam existir outras sequências consensos nestes promotores que ainda não foram estudadas, capazes de fornecer as mesmas características de flexibilidade das sequências consensos conhecidas dos elementos regulatórios.

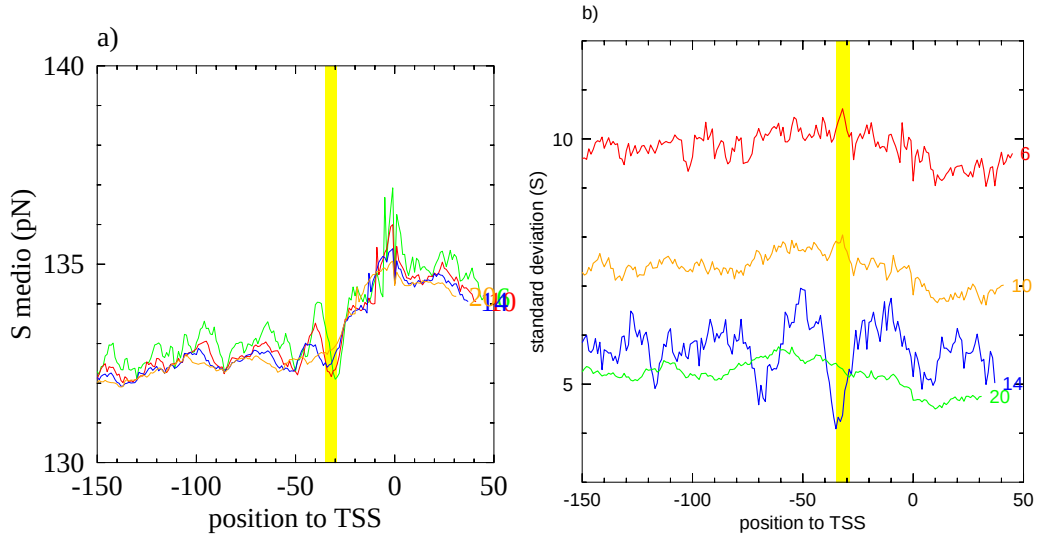


Figura 17

Perfil de flexibilidade média das sequências promotoras *"core-less"* de *H. sapiens* e o desvio padrão médio da flexibilidade das sequências promotoras reais *"core-less"* de *H. sapiens*. O cálculo da flexibilidade foi realizado com vários tamanhos de *n-mer*. Em ambos os gráficos, cada cor representa um tamanho de *n-mer*: vermelho (n=6), laranja (n=10), azul (n=14) e verde (n=20).

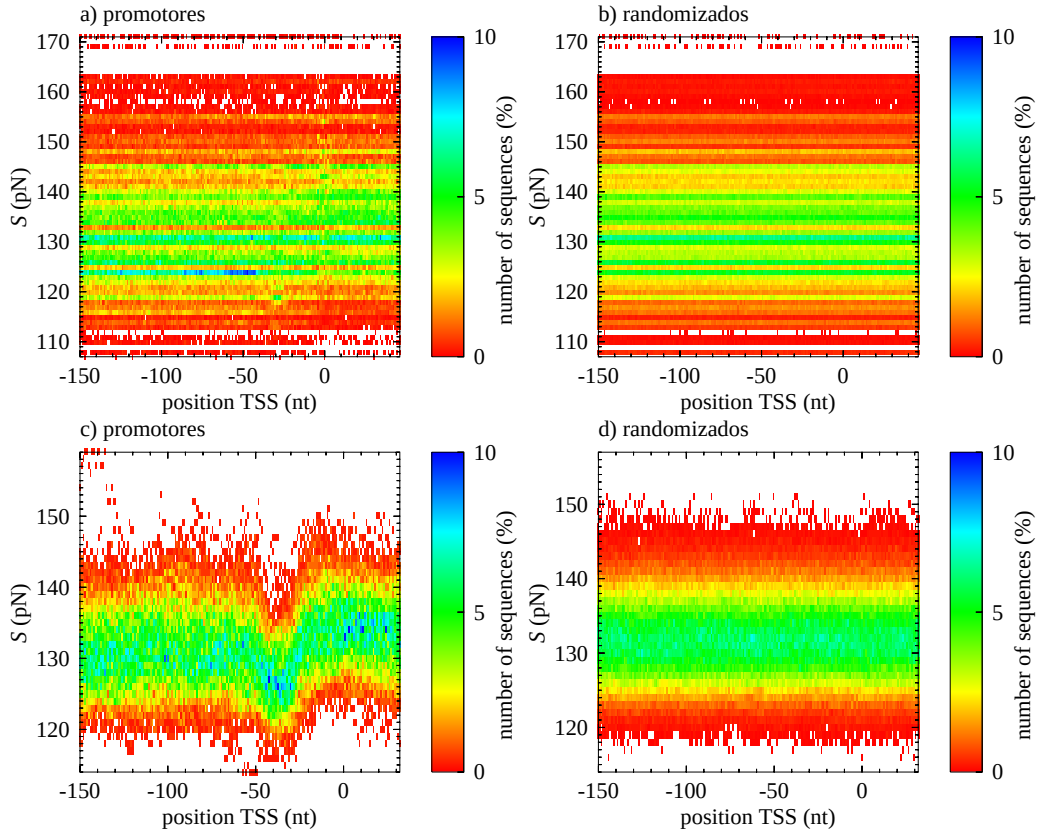


Figura 18
Mapas de cores das sequências promotoras *"core-less"* com *n-mer* de tamanho (a) 6 e (c) 20. (b) e (d) representam os mapas das sequências randomizadas *n-mer* de tamanho 6 e 20, respectivamente.

6.5 Caracterização termodinâmica da vizinhança dos microRNAs/mirtrons de *Drosophila melanogaster* e *Caenorhabditis elegans*

6.5.1 Introdução

Uma das principais motivações que fazem os cientistas procurarem por genes em um organismo é a incessante necessidade de entender como estes genes são regulados. Conforme vimos, as regiões promotoras dos genes contêm importantes elementos regulatórios e, portanto, estas regiões são capazes de controlar processos biológicos importantes tal como a síntese protéica. Além dos elementos regulatórios das regiões promotoras, na última década foi revelado que existem várias classes de RNAs não-codificandes que também podem funcionar como reguladores da síntese proteíca. Os microRNAs (ou miRNAs) é uma das classes mais estudadas pois estão envolvidos na regulação de numerosos processos celulares, incluindo a diferenciação, o desenvolvimento, a apoptose, a proliferação, a resposta ao stress e, além disso, eles também podem alterar a expressão de genes desencadeando várias doenças, tais como diabetes, câncer, e distrofia neuromuscular [75].

miRNAs são RNAs não-codificantes que foram identificados, pela primeira vez, no nematóide *Caenorhabditis elegans*, em 1 993, fruto de um gene chamado lin- 4 [76]. miRNAs canônicos são derivados de transcritos primários de miRNA (pri- miRNA), os quais são longas sequências de nucleotídeos que formam estruturas secundárias específicas em forma de hairpins. Pri- miRNA podem originar um ou mais hairpins, tipicamente com 55–70 nucleotídeos (nt) de comprimento. Em animais, os pri-miRNAs são clivados pela enzima nuclear Drosha RNase III para liberar os hairpins precursores dos miRNAs (pre-miRNA). Estes são então transportados para o citoplasma através da exportina-5 (Exp- 5) e lá, são clivados pela enzima Dicer RNase III para gerar um pequeno duplexo de miRNA/miRNA* [77]. Uma das fitas deste duplexo, chamada miRNA maduro (22–25 nt), é incorporado ao complexo RISC (complexo de silenciamento induzido pelo RNA) e guia o complexo ao seu alvo no mRNA para regular a expressão do gene enquanto a outra fita é rapidamente degradada [77, 78]. Em animais, a maioria das funções dos miRNAs estão relacionados com a regulação negativa dos genes, como por exemplo, as transcrições de miRNA do gene lin- 4 em *C. elegans*.

Ruby *et al.* [79] mostraram a existência de miRNAs intrônicas em *Drosophila melanogaster* e *C. elegans* que não passam pelo processamento da Drosha, proporcionando uma via alternativa para a biogênese miRNA [80]. Estes miRNAs foram chamados de "mirtrons" e a principal diferença entre eles e os miRNAs canônicos é que as seqüências

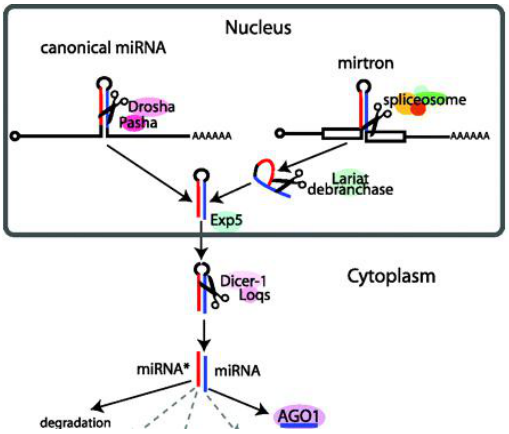


Figura 19
Biogênese dos microRNAs canônicos (à esquerda) e mirtrons (à direita). Figura retirada do artigo [83].

intrônicas formam laços e os mirtrons são originados pelo mecanismo de *splicing* [79–81]. Flynt *et al.* [82] publicou um trabalho onde ele reclassificou um subconjunto de mirtrons em *D. melanogaster* como "tailed mirtrons", os quais têm substanciais salências na porção 3' e são alvos do exossomo 3'-5', permitindo que miRNA pré-funcional seja gerado. A existência de mirtrons em mamíferos foi relatado por Berezikov *et al.* [81] onde, usando estratégias computacionais e experimentais, eles identificaram 3 mirtrons bem conservados que são expressos em diversos mamíferos, 16 mirtrons que são específicos de primatas e 46 candidatos à mirtrons em primatas.

Para mRNA, que é processado por *splicing*, Zhang *et al.* [84] verificaram que existe uma característica de conteúdo CG em torno dos sítios de *splice*. Foi mostrado que os *splicing* alternativos são promovidos pelas estruturas secundárias que se formam na fita de RNA [85], e que estas estruturas são fortemente determinadas pelo teor de CG [86]. MicroRNAs são co- expressados com mRNAs [87, 88] e, em particular, os mirtrons, aparentemente, não são processados pelo microprocessador Drosha mas por *splicing*. Como o *splicing* parece ser dependente da estabilidade termodinâmica, a pergunta que nos fazemos é a seguinte: poderia haver alguma característica de conteúdo CG que separa mirtrons de miRNAs comuns? De especial interesse, provavelmente são estas propriedades que suportam o mecanismo de *splicing* proposto para mirtrons [80].

Aqui nós nos propusemos a caracterizar sequências precursoras de miRNAs e mirtrons em termos de conteúdo CG e também de energia livre de Gibbs para *D. melanogaster* e *C. elegans*. Nós realizamos uma análise da termodinâmica destas sequências, buscando entender qual é o comportamento dos miRNAs em relação à sua vizinhança que poderia nos auxiliar na busca por promotores de genes que codificam miRNAs e mirtrons.

Nós encontramos que o conteúdo GC possui uma visível diferença em ambos os tipos de RNA de pequenas dimensões. Além disso, realizamos a mesma análise para mirtrons de mamíferos relatado por Berezikov *et al.* [81] e, mais uma vez, nossos resultados mostram diferenças importantes quando comparado com aqueles encontrados para os invertebrados.

6.5.2 Métodos

Para caracterizar os pequenos RNAs, nós comparamos o conteúdo CG das sequências precursoras que originam o miRNAs e mirtrons com o conteúdo CG de suas regiões vizinhas. A justificativa para essa abordagem é que, se a vizinhança da sequência de DNA tem uma diferença de estabilidade termodinâmica importante em comparação com a sequência precursora, logo, a fração de conteúdo CG destas duas porções também deve ser diferente. Para completar a nossa análise também calculamos as energias livre de Gibbs de mirtrons e miRNA comuns. Definimos a fração de CG conteúdo como

$$f = \frac{\text{número de nucleotídeos C e G}}{\text{número total de nucleotídeos}} \tag{10}$$

Dois tipos de razões de conteúdo CG foram usados, um relacionado com o precursor de miRNA ou mirtron f_P e o outro, f_N, se refere ao total de conteúdo CG dos 150 pares de bases a montante e a jusante da sequência precursora, que forma a vizinhança do precursor. Este comprimento de sequência flanqueadora foi o menor tamanho que leva a uma clara discriminação entre mirtrons e miRNAs canônicos. Ambos os conteúdos de CGs são combinados para formar uma razão entre o precursor e a sua vizinhança

$$R = \frac{f_P}{f_N} \tag{11}$$

Se o valor da razão é maior do que um ($R > 1$), o conteúdo CG f_P da sequência de precursora é maior do que a de seus vizinhos. Uma vez que o conteúdo CG está relacionado com a estabilidade termodinâmica, podemos inferir que $R > 1$ significa, geralmente, que a região flanqueadora do DNA é menos estável que a região precursora. Para facilitar, uasamos a notação

$$\begin{aligned} R^+ \rightarrow R > 1 \quad & \text{região } \textit{precursora} \text{ possui maior conteúdo CG}, f_P > f_N \\ R^- \rightarrow R < 1 \quad & \text{região } \textit{flanqueadora} \text{ possui maior conteúdo CG}, f_N > f_P \end{aligned}$$

A base de dados utilizada para obter os miRNAs precursores e mirtrons de invertebrados foi o mirBASE [89, 90], que é o principal repositório on-line de sequências de

microRNA. Os mirtrons relatados por Berezikov *et al.* [81] foram coletados dos dados complementares e as sequências vizinhas de cada mirtron foram obtidos do bancos de dados Ensemble API [91]. Para extrair as sequências que flanqueiam nós utilizamos o genoma completo de *D. melanogaster* versão r5.34 [92] e a versão WS223 para *C. elegans* [93]. Neste trabalho usamos o programa RNAfold do pacote de Vienna [94] com os parâmetros padrão para a obtenção das energias livre de Gibbs $\Delta G/N$ dos miRNAs precursoras e mirtrons em estudo.

6.5.3 Resultados e Discussão

Na fig. 20 mostramos a distribuição da razão do conteúdo CG (R) e a energia livre de Gibbs em função do tamanho de suas sequências precursoras $\langle \Delta G/N \rangle$ para miRNAs canônicos e mirtrons de *D. melanogaster*. Para miRNAs canônicos, a relação R de conteúdo CG (Fig. 20a) é quase uma distribuição gaussiana com centro em torno de $R = 1$. Isto significa que, para este tipo de miRNA, parece não haver relação preferencial do conteúdo CG dentro da sequência precursora e de seus vizinhos, apesar de um ligeiro desvio R^+ ser perceptível. Em contraste, todos os 18 mirtrons de *D. melanogaster* possuem $R < 1$ (R^-), como mostrado na fig. 20b. Mesmo sendo pequeno o número de mirtrons relatados, as chances de escolher 18 pequenos RNAs com R^- por acaso, considerando a distribuição de miRNA canônico, é menos de 10^{-6}. Uma possível explicação para a predominância de R^- seria se as regiões intrônicas tivessem alto teor de CG. No entanto, regiões intrônicas de *D. melanogaster* tem um dos menores conteúdo CG neste genoma: 0.4, em comparação com 0.52 para regiões codificantes. Portanto, podemos concluir que R^-, o qual significa que a região flanqueadora é mais estável do que a região precursor, pode possivelmente desempenhar um papel considerável na biogênese mirtron. Para miRNAs canônicos de *C. elegans* observa-se uma similar distribuição gaussiana em relação a R (Fig. 20c), mas com um viés muito mais forte para R^+. Esse viés também é refletida pela relação $n^+ : n^-$, mostrada na Tab. 5, isto é, há mais de 3 miRNAs R^+ para cada miRNA R^-. No entanto, apesar de existir, até o momento, apenas cinco mirtrons relatados, todos eles mostram R^-(Fig. 20d), semelhante ao *D. melanogaster*.

Dado que as mirtrons de *D. melanogaster* e *C. elegans* possuem uma distinta distribuição R^-, seria possível chegar a conclusões semelhantes através da análise da energia livre de Gibbs? Na fig. 21, mostramos a distribuição de energia livre de Gibbs, $\Delta G/N$, para ambos os invertebrados e as quantidades detalhadas também são mostrados na Tab. 5. Para os pre-miRNAs canônicos observamos, novamente, uma distribuição gaussiana nas Figs. 21a e 21c. No entanto, a distribuição para mirtrons está localizada em valores muito mais elevados de $\Delta G/N$, tipicamente acima de -30 kcal/mol, Figs. 21b e 21d. Em parte,

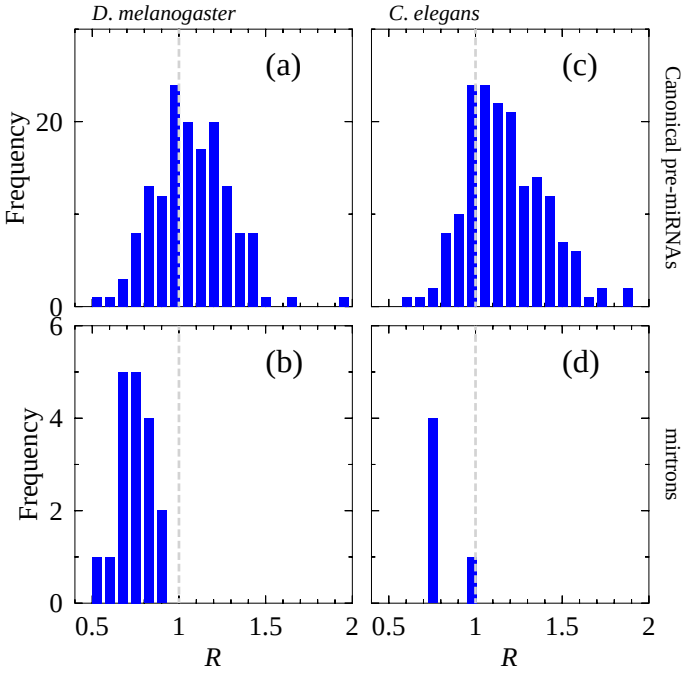


Figura 20
Distribuição da razão de conteúdo CG (R), para a) 151 pre-miRNAs canônicos e b) 18 mirtrons de *D. melanogaster* e c) 170 miRNAs canônicos e d) 5 mirtrons de *C. elegans*.

Organismo	tipo de RNA	total	n^+	n^-	$n^+ : n^-$	$\langle R \rangle$	$\langle \Delta G/N \rangle$ (kcal/mol∗nt)
C. elegans	mirtrons	5	0	5		0.81 ± 0.08	-0.33 ± 0.09
D. melanogaster	mirtrons	18	0	18		0.76 ± 0.10	-0.32 ± 0.05
D. melanogaster	tailed mirtrons	7	2	5	1 : 1.8	0.84 ± 0.24	-0.21 ± 0.06
C. elegans	miRNAs canônicos[1]	170	131	39	3.3 : 1	1.18 ± 0.23	-0.38 ± 0.09
D. melanogaster	miRNAs canônicos[1]	151	97	54	1.8 : 1	1.08 ± 0.21	-0.36 ± 0.06

[1] mirtrons excluídos.

Tabela 5
Características de conteúdo CG e energia livre dos pre-miRNAs canônicos e dos mirtrons de invertebrados. Também são mostrados o número de sequências $n^{\pm}$ com $R^{\pm}$, média da razão de conteúdo CG $\langle R \rangle$, média da energia livre $\langle \Delta G \rangle$ pelo comprimento médio do precursor $\langle N \rangle$.

os resultados mais altos de $\Delta G/N$ se deve ao fato de que os mirtrons de invertebrados são consideravelmente mais curtos do que os miRNAs. Existe algum tipo de associação entre mirtrons com maior $\Delta G/N$ com aqueles que são R^-? Para descobrir, nós isolamos todos os miRNAs canônicos com R^- e recalculamos a distribuição de energia livre como é mostrado nas barras vermelhas das Figs. 21 a e 21 c. Sem dúvida, os miRNAs com R^- tendem a ter valores de $\Delta G/N$ um pouco mais elevados, ver também Tab. 6. Isto não é surpreendente, dado que o R^- implica em precursor com menor conteúdo CG e, portanto, termodinamicamente menos estável. No entanto, a análise da distribuição de energia livre é muito menos conclusiva em termos de diferenças entre mirtrons e miRNAs canônicos do que a razão do conteúdo GC.

As próximas questões é saber se outros tipos de mirtrons relatados, tais como mirtrons de primatas e de mamíferos mostram a mesma distribuição R apresentada para os

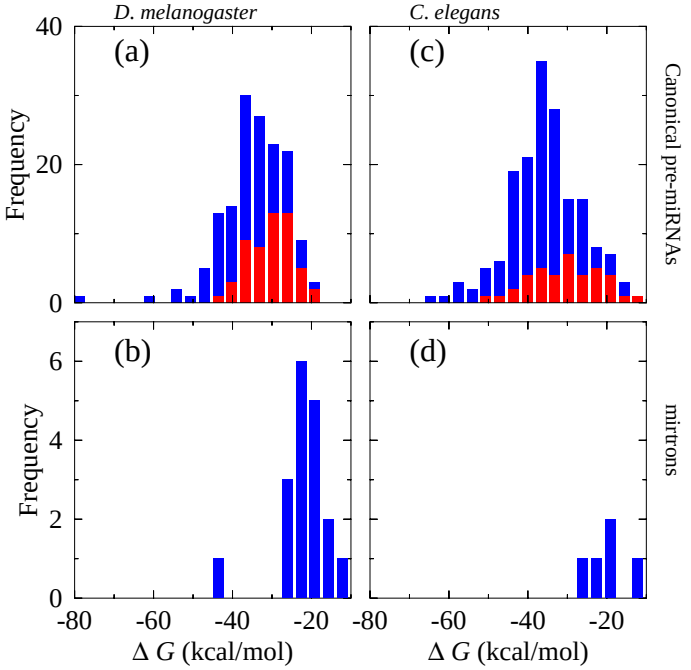


Figura 21
Distribuição da energia livre de Gibbs para a) 151 pre-miRNAs canônicos e b) 18 mirtrons de *D. melanogaster*. e c) 170 pre-miRNAs canônicos e d) 5 mirtrons de *C. elegans*. As barras vermelhas representam os pre-miRNAs com R^-.

invertebrados. Como mostrado na Tab. 7, em termos de teor de CG, energia livre e comprimento, estes mirtrons estão muito mais próximos de miRNAs canônicos. Na verdade, eles tendem geralmente, para R^+, e, de forma consistente, para $\Delta G/N$ mais negativo, ver Figs. 22 e 23.

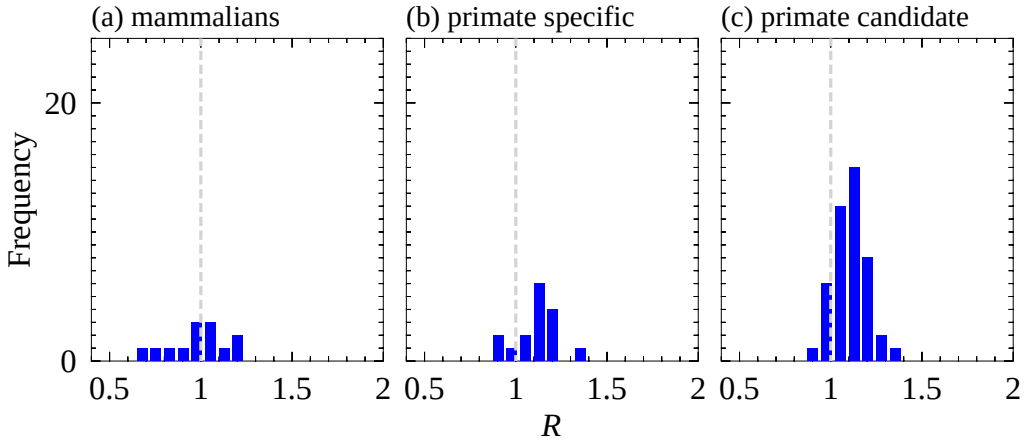


Figura 22
Distribuição do conteúdo CG para a) 13 prováveis mirtrons de mamíferos, b) 16 mirtrons específicos de primatas e c) 46 candidatos a mirtrons de primatas.

.

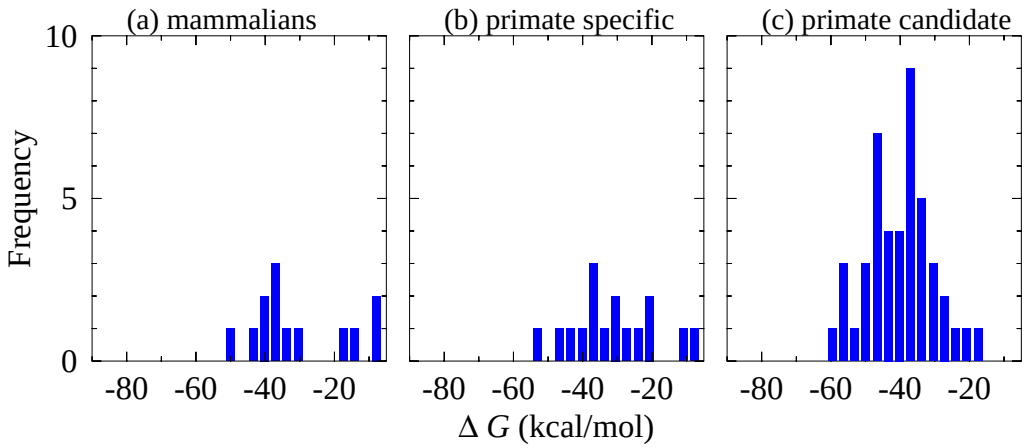


Figura 23
Distribuição da energia livre de Gibbs para a) 13 prováveis mirtrons de mamíferos, b) 16 mirtrons específicos de primatas e c) 46 candidatos a mirtrons de primatas.

Organismo	tipo de RNA	n^-	$\langle R^- \rangle$	$\langle \Delta G/N \rangle$ (kcal/mol∗nt)
C. elegans	miRNAs canônicos	39	0.90 ± 0.08	-0.33 ± 0.08
D. melanogaster	miRNAs canônicos	54	0.87 ± 0.10	-0.33 ± 0.04

Tabela 6
Características de conteúdo CG e energia livre dos pre-miRNAs canônicos considerando apenas aqueles com R^-.

Organismo	tipo de RNA	total	n^+	n^-	$n^+ : n^-$	$\langle R \rangle$	$\langle \Delta G/N \rangle$ (kcal/mol∗nt)
mamíferos	prováveis mirtrons	13	8	5	1.5 : 1	0.98 ± 0.16	-0.34 ± 0.15
primatas	específicos mirtrons	16	13	3	4.3 : 1	1.12 ± 0.11	-0.38 ± 0.13
primatas	candidatos à mirtrons	45[2]	40	5	8 : 1	1.11 ± 0.09	-0.45 ± 0.09

[2]Ref. [81] relata 46 candidatos à mirtrons para primatas mas o material suplementar traz 45 sequências.

Tabela 7
Características de conteúdo CG e energia livre dos mirtrons específicos de vertebrados.

7 Conclusão

Para alcançar o objetivo geral e os objetivos específicos deste projeto (listados na seção 4), foi necessário dividir o nosso trabalho em 5 sub-projetos cujos resultados foram apresentados na seção 6. Os resultados obtidos com os 4 sub-projetos iniciais nos forneceu dados suficientes para podermos afirmar que, além de ser a maneira fisicamente correta de predizer a flexibilidade de um promotor, o nosso modelo de soma inversa das constantes elásticas de cada *n-mer* resulta em conclusões que são condizentes com sentido biológico da transcrição: a porção *downstream* ao TSS apresenta-se mais rígida do que a porção *usptream* ao TATA-*box* e a posição em torno de -28 é uma das regiões mais flexíveis.

De acordo com o que descrevemos na introdução, o processo transcricional é uma etapa da síntese protéica que depende do reconhecimento e da abertura de certas porções da sequência de DNA, conhecidas como elementos regulatórios. O TATA-*box*, por exemplo, é um dos elementos regulatórios mais estudados e sua interação com a proteína TBP facilita que esta região seja desenrolada. Estudos realizados com o TATA-*box* [71, 95, 96] relatam que este elemento apresenta uma tendência a ser flexível devido ao fato de se deformar quando interage com a proteína e os nossos resultados do perfil de flexibilidade confirmam esta tendência. Através da soma direta, Zeng [11] concluiu em seu trabalho que esta região apresenta-se mais rígida, o que contradiz com os estudos realizados e com a lógica da transcrição.

Nós concluímos também que organização e a manutenção das bases presentes nas regiões promotoras são de extrema importância para que as propriedades físicas possam ser interpretadas pelos complexos enzimáticos. Em nossas análises envolvendo as sequências promotoras randomizadas dos organismos do banco de dados EPD nós vimos que a flexibilidade do DNA apresenta-se bastante uniforme, ao contrário das sequências promotoras.

O cálculo da flexibilidade média nos trouxe informações gerais bastante interessantes mas, este tipo de medida, por si só, não representa muito bem aqueles promotores que se dispersam muito do valor médio. Utilizá-la como parâmetro de busca de promotores não é uma medida segura quando o desvio padrão médio da flexibilidade apresenta-se relativamente alto e parecido com os valores encontrados para o desvio padrão médio da flexibilidade das sequências promotoras randomizadas. Diante disso, o estudo da frequência de cada valor de flexibilidade produz dados mais completos sobre o comportamento da flexibilidade em sequências promotoras reais.

A utilização dos mapas de cores para visualizar as frequências da flexibilidade por

posição facilita a análise das diferenças entre as sequências promotoras e as randomizadas, permitindo avaliar seu potencial para predição de promotores. Através delas, foi possível ver o que é igual entre as sequências promotoras reais e randomizadas de cada organimo mas, o principal foi enxergar as diferenças. São as diferenças geradas por cada *n-mer* que são importantes para distinguir uma sequência com potencial para ser promotor de uma que não é e são nelas que precisamos focar. Além disto, fica bastante evidente que usar *n-mer* de tamanho pequeno, 6 ou 7, sob o argumento de que são do mesmo tamanho dos elementos regulatórios não se justifica já que as diferenças são muito mais acentuadas para *n-mer* de tamanhos maiores.

No último sub-projeto apresentado, nós também introduzimos o conceito de proporção do conteúdo CG das seqüências precursoras e regiões flanqueadoras dos genes de microRNAs e mirtrons e mostramos que eles fornecem um método de caracterização para mirtrons de invertebrados. Já para mirtrons de mamíferos, parece existir uma tendência para um maior conteúdo CG na região precursora. Em ambos os casos encontramos uma clara assimetria na distribuição de conteúdo CG. Esta assimetria, quando comparado ao miRNA normal, parece apoiar a noção de que mirtrons são processadas de um modo semelhante ao mRNA, em vez de ser processado pela Drosha.

8 Perspectivas futuras

Os resultados do nosso projeto evidenciaram que para a predição de promotores vale a pena investir no estudo sistemático dos parâmetros antes de partir para a busca propriamente dita. Mas para responder a pergunta sobre a eficácia real dos parâmetros na predição temos que implementar um método de busca de promotores usando as probabilidades respresentadas pelos mapas de cor, como por exemplo usando classificadores Bayesianos.

Vemos que embora promissora, a flexibilidade do DNA pode não ser suficiente para uma predição robusta. Podemos nos fazer então a seguinte pergunta: quais outros parâmetros podem ser melhores para esta tarefa? Ao invés de selecionar parâmetros que representam aspectos físicos do DNA, como de flexibilidade, porque não podemos usar parâmetros otimizados para busca de promotores? Neste sentido, pensamos em usar métodos de otimização por algorítmos genéticos para buscar parâmetros que maximizem a diferença dos mapas de probabilidade entre promotores e sequências randomizadas.

Finalmente, uma vez de posse de um bom método de predição de promotores, pretendemos adaptar este método para predição de promotores de microRNA unindo assim as duas linhas de pesquisa que foram objeto desta dissertação.

A Mapas de cores adicionais com *n-mer* igual à 10 e 14

Abaixo estão representados os mapas de cores com *n- mer* igual à 10 e 14 para os promotores de *Homo sapiens*, *Drosophila melanogaster*, *Oriza sativa* e *"core-less"* de *H. sapiens*:

A.1 *Homo sapiens*

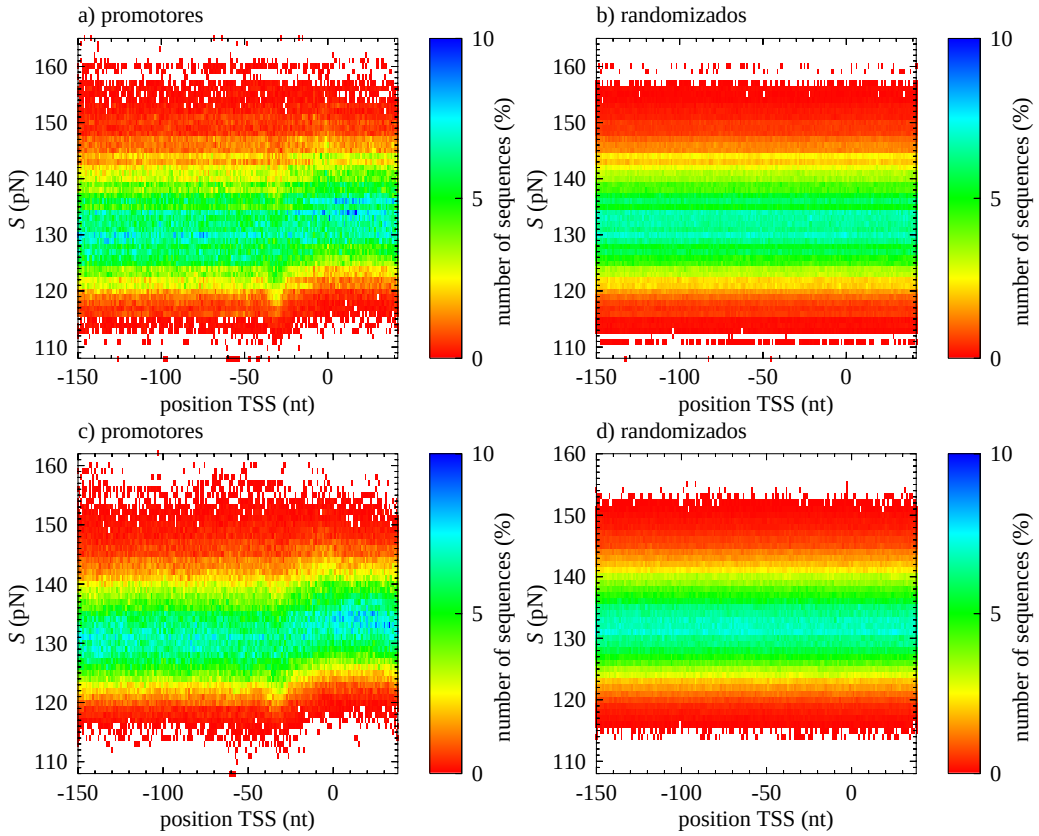


Figura 24
Mapas de cores das sequências promotoras de *H. sapiens* com *n-mer* igual à (a) 10 e (c) 14. (b) e (d) representam os mapas das sequências randomizadas *n-mer* de tamanho 10 e 14, respectivamente.

A.2 *Drosophila melanogaster*

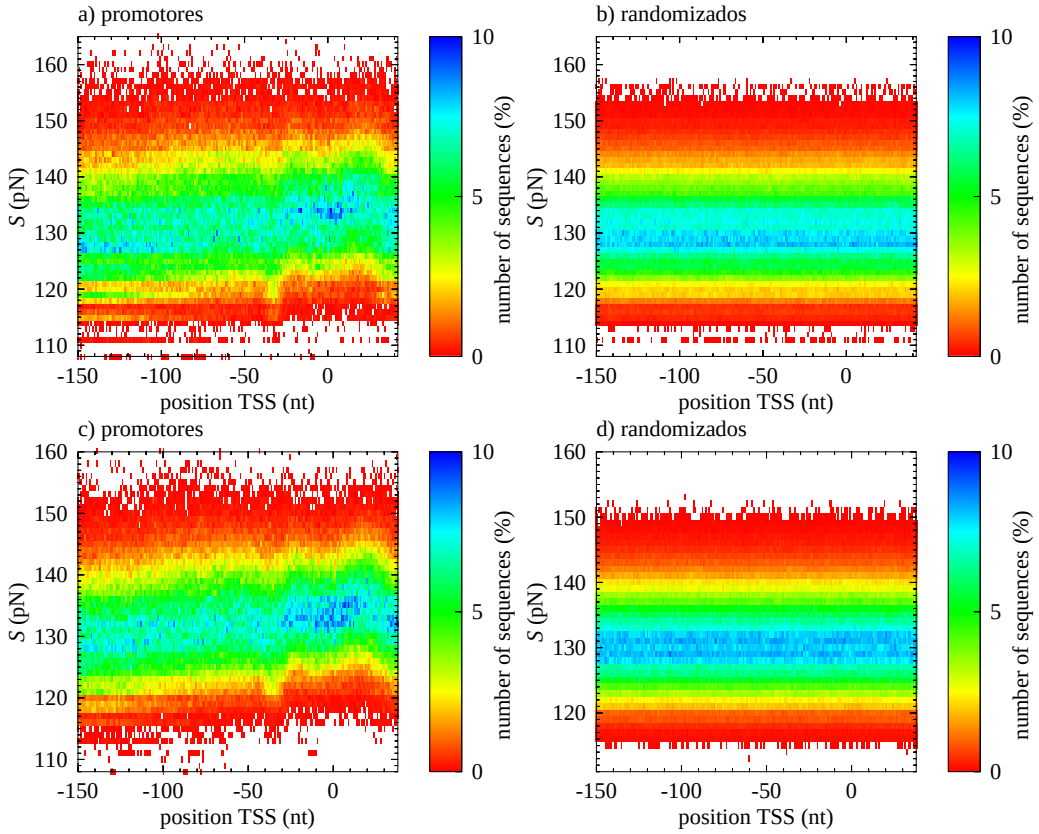


Figura 25
Mapas de cores das sequências promotoras de *D. melanogaster* com *n-mer* igual à (a) 10 e (c) 14. (b) e (d) representam os mapas das sequências randomizadas *n-mer* de tamanho 10 e 14, respectivamente.

A.3 *Oriza sativa*

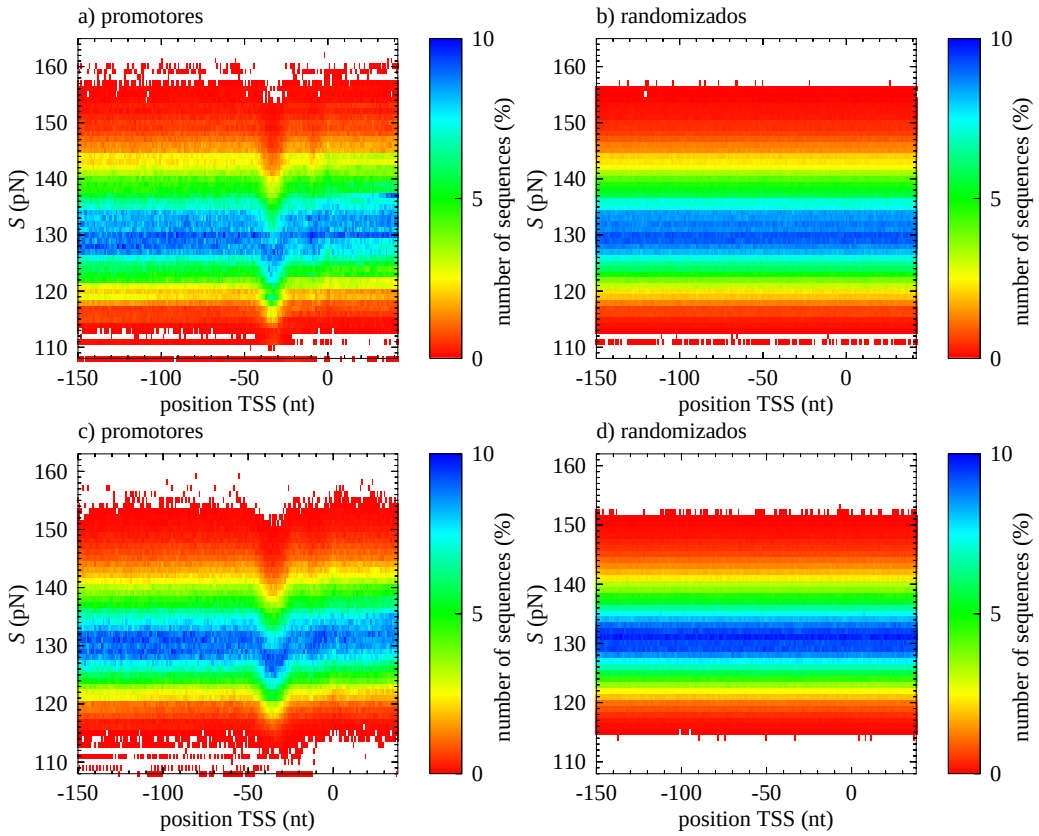


Figura 26
Mapas de cores das sequências promotoras de *O. sativa* com *n-mer* igual à (a) 10 e (c) 14. (b) e (d) representam os mapas das sequências randomizadas *n-mer* de tamanho 10 e 14, respectivamente.

A.4 *"Core-less"* de *Homo sapiens*

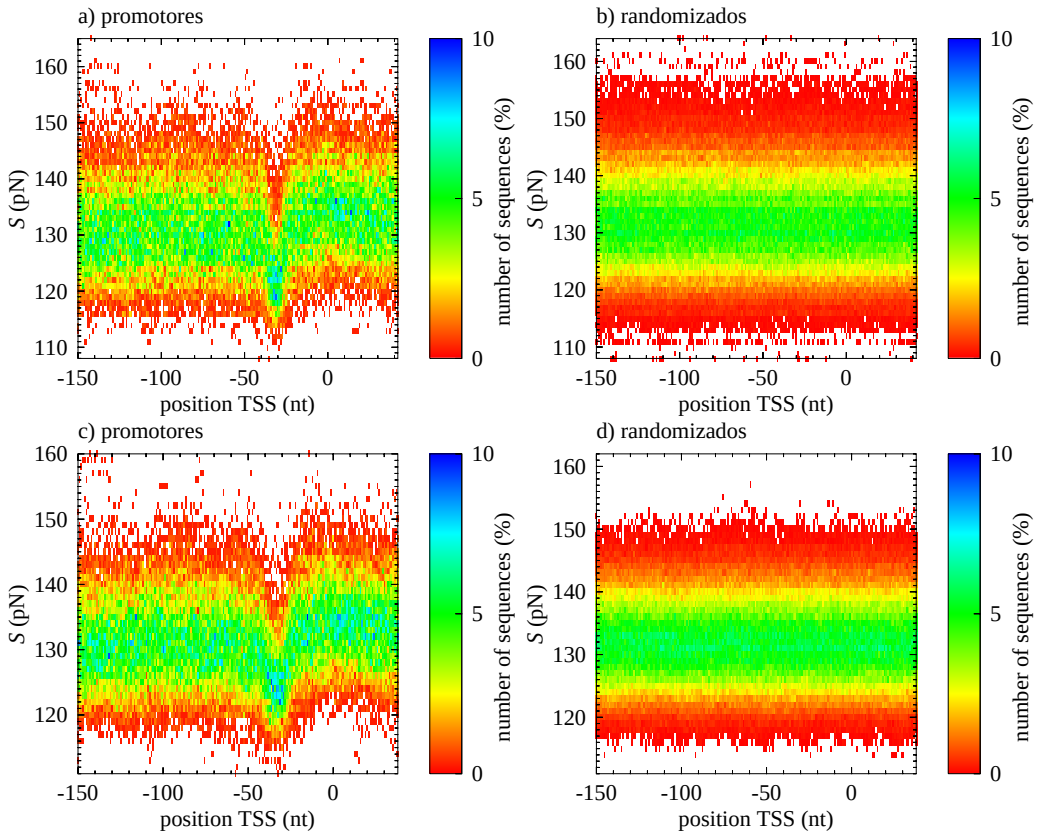


Figura 27
Mapas de cores das sequências promotoras de *"core-less"* com *n-mer* igual à (a) 10 e (c) 14. (b) e (d) representam os mapas das sequências randomizadas *n-mer* de tamanho 10 e 14, respectivamente.

B Scripts utilizados neste trabalho

B.1 media.pl

Este programa calcula o valor de flexibilidade média em cada posição das sequências promotoras, de acordo com o tamanho de *n-mer* estabelecido.

```
use List::Util qw(first max maxstr min minstr reduce shuffle sum);
use lib '/home/denise/Mestrado/modulos';
use subseq;


if (scalar(@ARGV) < 3)
   {
   print "\n\nUso: perl flexpromoterperfil.pl <nucleotideos> <saida>
   <entrada1> \n\n";
   exit;
   }

my $nn=$ARGV[0];
my $saida=$ARGV[1];
my $arquivo=$ARGV[2];
# my $arquivo2;
# if (exists $ARGV[3]) {$arquivo2=$ARGV[3];}
# else                 {$arquivo2=$arquivo;}
print "arquivo de saida $saida$nn.histo\n";

open(SAID,">$saida$nn.histo") or die "Nao pude abrir $saida$nn.histo\n";
#open(OI,">$nn-$nn.dat");
open(ARQ, $arquivo) or die "Nao pude abrir $arquivo\n";

@seqs=<ARQ>;
# foreach $linha(@seqs)
# {if ($linha=~/N/)
# {print $linha;}}

open (K,'stat22-69he.par') or die 'Nao pude abrir o arquivo';
 while ($linha = <K>)
   {
     if ($linha =~ /(.).\_(.).:harmonic\.k (.*)/) # le apenas os valores de
                                                  # flexibilidade
       { $elastic{"$1$2"}=$3; }
   }
   $elastic{"TT"}=$elastic{"AA"};
   $elastic{"TG"}=$elastic{"CA"};
   $elastic{"GG"}=$elastic{"CC"};
   $elastic{"CT"}=$elastic{"AG"};
   $elastic{"GT"}=$elastic{"AC"};
   $elastic{"TC"}=$elastic{"GA"};
#exit;

# print scalar(%elastic);
# for my $k (sort keys %elastic)
#   {
#   print "$k $elastic{$k}\n";
#   }
# exit;

$cont=0;
$contseq=0;
@flex=();# array com kmrdio
@desvio=();
```

```
  foreach $linha(@seqs)
    {
    chomp($linha); #este chomp aqui retira a mudanca de linha
    if (not $linha =~ />.*/)
      {
      @n_uplos=nmer($linha,$nn);
      #print print "[",join('||',@n_uplos),"]\n\n";
      $contseq++;
       #print @n_uplos;

      $cont=0;
      foreach $sext (@n_uplos) # pega cada trio da lista trios e faz o corte em
                               # 2 duplas pra somar os valores de k e ter o
                               # kequivalente
        {
        my $keq=keq($sext,$nn,\%elastic);
       # print $keq;
        $flex[$cont]+=$keq; # esta parte pega todas os valores de keq de cada
                            # posicao e vai somando um valor em cima do outro.
        $cont=$cont+1;      # este $cont faz a contagem de quantas
                            # posicoes tem as sequencias que e 196
        #print $keq;
        }
     }
   # last if ($contseq == 100);
    }

  #print "$contseq sequencias no arquivo\n";
  for($cont=0; $cont < scalar(@flex); $cont++)
     {
   print SAID  "$cont ",$flex[$cont]/$contseq,"\n"
     }
```

B.2 desvio.pl

Este programa calcula o valor de desvio padrão médio da flexibilidade em cada posição das sequências promotoras, de acordo com o tamanho de *n-mer* estabelecido.

```
  use List::Util qw(first max maxstr min minstr reduce shuffle sum);
  use lib '/home/denise/Mestrado/modulos';
  use subseq;

  if (scalar(@ARGV) < 3)
    {
    print "\n\nUso: perl fpromoterperfil.pl <nucleotideos> <saida>
                        <entrada1> <entrada2>\n\n";
    exit;
    }

  my $nn=$ARGV[0];
  my $saida=$ARGV[1];
  my $arquivo=$ARGV[2];
  my $arquivo2;

  if (exists $ARGV[3]) {$arquivo2=$ARGV[3];}
  else                 {$arquivo2=$arquivo;}

  print "arquivo de saida $saida$nn.histo\n";
  open(SAID,">$saida$nn.histo") or die "Nao pude abrir $saida$nn.histo\n";


  open(ARQ, $arquivo) or die "Nao pude abrir $arquivo\n";
  @seqs=<ARQ>;
  # foreach $linha(@seqs)
```

```
# {if ($linha=~/N/)
# {print $linha;}}


#open (K,'stat22-1020he.par') or die 'Nao pude abrir o arquivo'; # arquivo
                                                                 #para teste
open (K,'stat22-69he.par') or die 'Nao pude abrir o arquivo';
 while ($linha = <K>)
  {
    if ($linha =~ /(.).\_(.).:harmonic\.k (.*)/)  # le apenas os valores
                                                  # de flexibilidade
      { $elastic{"$1$2"}=$3; }

  }
  $elastic{"TT"}=$elastic{"AA"};
  $elastic{"TG"}=$elastic{"CA"};
  $elastic{"GG"}=$elastic{"CC"};
  $elastic{"CT"}=$elastic{"AG"};
  $elastic{"GT"}=$elastic{"AC"};
  $elastic{"TC"}=$elastic{"GA"};

$cont=0;
$contseq=0;
@flex=();# array com kmedio
@desvio=();

foreach $linha(@seqs)
  {
   chomp($linha);
  if (not $linha =~ />.*/)
    {
    @n_uplos=nmer($linha,$nn);
    $contseq++;

    #print @quadruplos[197];
    #$pos=-150;#onde posicao inicial de cada sequencia

    $cont=0;
    foreach $sext (@n_uplos) # pega cada trio da lista trios e faz o corte em
                             # 2 duplas pra somar os valores de k e ter o
                             # kequivalente
      {
      my $keq=keq($sext,$nn,\%elastic); #KEQ OU KEQSD??

      #print $cont;
      $flex[$cont]+=$keq; # esta parte pega todas os valores de keq de
                          # cada posicao e vai somando um valor em cima do outro.
      $cont=$cont+1; # este $cont faz a contagem de quantas
                     # posicoes tem as sequencias que e 196 nt
      #print $keq;

      }
    }
  }

#print "$contseq sequencias no arquivo\n";

# for($cont=0; $cont < scalar(@flex); $cont++)
#    {
#     $flex[$cont] /= $contseq; # fazendo a media
#   "$cont ",$flex[$cont],"\n";
#    print $cont,$flex[$cont];}

for($cont=0; $cont < scalar(@flex); $cont++)
   {$flex[$cont]= $flex[$cont]/$contseq;
    print "$cont ",$flex[$cont],"\n"
   }
```

```
96
   open(ARQ, $arquivo2) or die "Nao pude abrir $arquivo2\n";
   @seqs=<ARQ>;

   $contseq=0;
101 foreach $linha(@seqs)
     {chomp($linha);
     if (not $linha =~ />.*/)
       {
       @n_uplos=nmer($linha,$nn);
106    $contseq++;


       #print @quadruplos[197];

111    #$pos=-150;#onde posicao inicial de cada sequencia

       $cont=0;
       foreach $sext (@n_uplos) # pega cada trio da lista trios e f
                                # faz o corte em 2 duplas pra somar os
116                             # valores de k e ter o kequivalente
         {
         my $keq=keq($sext,$nn,\%elastic); #inicialmente o valor de kequivalente e 0

         $desvio[$cont] += ($flex[$cont] - $keq )**2;
121      $cont=$cont+1;
        #$desvio[$cont] += (($flex[$cont] - $keq)/$flex[$cont] )**2;
         #$desvio[$cont] += abs(($flex[$cont] - $keq) /$flex[$cont]);
         }
       }
126  }

   #ABAIXO, TRECHO PARA DESVIO  DE SOMA INVERSA
   ##############################################
   my $desviototal=0;
131 for($cont=0; $cont < scalar(@flex); $cont++)
      {
     $desvio[$cont] = sqrt($desvio[$cont]/$contseq);
     $desviototal += $desvio[$cont];
   ## print SAID "$cont ",$flex[$cont]-$desvio[$cont]," ",$flex[$cont]," ",
136  $flex[$cont]+$desvio[$cont], "\n"; #para fazer grafico com valor de media mais
                                        #desvio
     print SAID "$cont ",$desvio[$cont], "\n"; #pra fazer desvio para soma inversa
   ## print SAI "$cont ",$desvio[$cont]/$contseq, "\n"
      }
141

   print "arquivo1=$arquivo arquivo2=$arquivo2\n";
   print "desvio padrao por nt = ",$desviototal/scalar(@flex), "\n";
   ##############################################
```

B.3 random.pl

Este programa gera os arquivos com as sequência promotoras randomizadas.

```
use List::Util qw(first max maxstr min minstr reduce shuffle sum);
use lib '/home/denise/Mestrado/modulos';
use subseq;

if (scalar(@ARGV) < 3)
   {
   print "\n\nUso: perl fpromotersprelativo.pl <nucleotideos> <saida>
   <entrada1> \n\n";
   exit;
   }

my $nn=$ARGV[0];

my $saida=$ARGV[1];

my $arquivo1=$ARGV[2]; #concatenadohs.dat

print "arquivo de saida $saida$nn.dat\n";
open(SAI,">$saida$nn.dat");

open(ARQ, $arquivo1) or die "Nao pude abrir $arquivo1\n";
@seqs=<ARQ>;
print scalar(@seqs)," linhas no arquivo $arquivo1\n";
# foreach $linha(@seqs)
#  {if ($linha=~/N/)
#  {print $linha;}}



open (K,'stat22-69he.par') or die 'Nao pude abrir o arquivo';
 while ($linha = <K>)
   {
     if ($linha =~ /(.).\_(.).:harmonic\.k (.*)/)  # le apenas os valores de
                                                   # flexibilidade
       { $elastic{"$1$2"}=$3; }
   }

   $elastic{"TT"}=$elastic{"AA"};
   $elastic{"TG"}=$elastic{"CA"};
   $elastic{"GG"}=$elastic{"CC"};
   $elastic{"CT"}=$elastic{"AG"};
   $elastic{"GT"}=$elastic{"AC"};
   $elastic{"TC"}=$elastic{"GA"};

my @seqr=@seqs;
for (my $i=0; $i < 3; $i++) {push(@seqr,@seqr);} #crio uma pilha de sequencias
                                                 #randomizadas, apartir das
                                                 #sequencias reais



my ($minr,$maxr,$nupr,@spr)=sp(\@seqr,$nn,\%elastic,1); # 4o argumento,
                                                        # faz a randomizacao
                                                        # e faz a normalizacao
                                                        # dos valores de
                                                        # estiramento encontrados.
print "Randomizado ",scalar(@seqr)," min=$minr max=$maxr n-nuplos=$nupr\n"; #nos
                                          #da o numero de sequencias randomizadas,
                                          #o valor minimo e maximo de estiramento
                                          #e o numero de posicoes das seqs
                                          #randomizadas.
print_sp($minr,$maxr,$nupr,\@spr,"sp$nn-r.dat"); # essa parte gera o sprel3-r, so
                                          #com valores de estiramento normalizados
```

```
                                        #(n e a relacao)

print scalar(@spr)," ",scalar(@{$spr[scalar(@spr)-1]}),"\n";


my ($mins,$maxs,$nups,@sp)=sp(\@seqs,$nn,\%elastic,0);# faz a normalizacao dos
                                                       # valores de estiramento
                                                       # encontrados.
print "Normal min=$mins max=$maxs n-nuplos=$nups\n"; #nos da o valor minimo e
                                                     #maximo de estiramento
                                                     #e o numero de posicoes
                                                     #das seqs reais.
print_sp($mins,$maxs,$nups,\@sp,"sp$nn-s.dat");# essa parte gera o sprel3-s, so com
                                               #valores de estiramento normalizados
                                               #(n e a relacao)


print scalar(@sp)," ",scalar(@{$sp[scalar(@sp)-1]}),"\n";

my $linhas=($maxs-$mins+1);# essa informacao nos da a variacao
                           # dos valores.Sao uteis para a
                           # confeccao dos graficos de cor.
                           #Esta numa subrotina print_sp.
print SAI "$mins $maxs $linhas $nups\n";


for (my $i=$mins; $i<=$maxs; $i++) #essa parte gera o arquivo sp3.dat(e a relacao)
  {
  for (my $j=0; $j<$nups; $j++)
   {
#      print SAI "$i $j $sp[$i][$j] \n"; #$i=valores de keq/ $j= posicoes /
                                         #$sp[$i][j]= quantas vezes, ou seja,
                                         #quantas vezes aparece tal valor
#                                        #de keq em determinada seq.
   my $rel;
   if (exists $sp[$i][$j] and exists $spr[$i][$j]) # aqui, os valores de
                                         # estiramento que entram para
                                         # obtermos a razao ja vem
                                         # normalizados la de cima.
     {
     $rel= $sp[$i][$j]/$spr[$i][$j];
     }
   else {$rel=0;} #para aqueles valores de estiramento que nao existem,
                  #atribuimos zero para essa razao.

#   $rel=0 if (($rel < 0.5) or ($rel < 4));
   print SAI  sprintf("%f ",$rel);
    }
    print SAI "\n";
 }
```

B.4 findelements.pl

Este programa gera o arquivo com promotores *"core-less"* de *H. sapiens*.

```
open(SAID, ">exttodos.dat") or die "Nao pode abrir o arquivo"; #seqs sem elementos
                                                               #regulatorios
open(SAIDA,">sotodos.dat") or die "Nao pude abrir o arquivo\n";#seqs so c elementos
                                                               #regulatorios


#open(ARQ, 'concatenadohs.dat') or die "Nao pude abrir o arquivo\n";

@seqs=<ARQ>;


@tata=(); #array com as sequencias que so tem elementos regulatorios conhecidos


foreach $s(@seqs) # para tamanho de sequencia igual a 201
 {
    $tata=substr($s,100,41); #TATA (-50 a -10)
    $inr=substr($s,140,21); #INR (-10 a +10)
    $dpe=substr($s,174,26); #DPE (+25 a +51)

    #print  $tata,"\n";

    if (
       # $tata=~/ATAT/
        not $tata=~/TATA[A,T]A[A,T]/      and not    #TATA
        $inr=~/[C,T][C,T]A[A,T,C,G][T,A][C,T][C,T]/ and not  #INR
        $dpe=~/[A,G]G[T,A]CGTG/                      #DPE

          )

     {push @tata,$s;}
    else
     {}

   }
print SAIDA @tata;
#print scalar (@tata);

for(my $i=0; $i < scalar(@tata); $i++) #elimina as sequencias com elementos
                                       #regulatorios do arquivo concatenadohs.dat

    { $busca=$tata[$i];
     #print SAIDA $busca;
    $pos=0;
    foreach  $linha (@seqs)
      {
       if ($linha=~ /$busca/)
        {
          splice(@seqs,$pos,1);
#          $cont++;
          $pos--;# como deletou uma linha do @promoter, todas as posicoes sobem 1
                 #posicao, entao e preciso voltar o contador da posicao 1 x.
        }
       $pos++; }
           }

 foreach (@seqs){ #imprimi cada elemento do array em uma linha ( $_)
  print SAID $_;
}
```

B.5 findhairpin.pl

Este programa faz a busca dos miRNAs/mirtrons nos genomas dos organimos e extrai as sequências já com a vizinhança que desejamos.

```
use POSIX qw(ceil);
$microrna="cbr140.dat.gz";
#$microrna="/share/bioinf/database/hairpin-dme.fa.gz";
#$microrna="../mirtron/mirtrons.fa.gz";
#$genome="c_elegans.WS223.dna.fa.gz";
$genome="c_briggsae.WS200.genomic.fa.gz";
#$genome="dpse-all-chromosome-r2.9.fasta.gz";
#$genome="dmel-3R-dna-fasta.dat.gz";

$limite=150;
$outputfile="mirtronsanalisados.dat";
open(VIZINHO,">cbrt-$limite.dat");

#$antes_outputfile="resultados antes.dat";
#open(ANTES,">antes-$outputfile");

#$depois_outputfile="resultados depois.dat";
#open (DEPOIS,">depois-$outputfile");

$lines_in_buffer=1000;


@hairpins_id;
@hairpins;
@r_hairpins;
@c_hairpins;
@rc_hairpins;

$largest_hairpin=0;

sub read_hairpins{
  open(HAIRPIN,$_[0]);
  my $linha="";
  $hairpin_count=-1;
  while (chomp($linha=<HAIRPIN>))
    {
    if ($linha =~ />.*/)
      {
      $hairpin_count++;
      $hairpins_id[$hairpin_count]=$linha;
      $hairpins[$hairpin_count]="";
      }
    else
      {
      $hairpins[$hairpin_count]="$hairpins[$hairpin_count]$linha";
      }
    }
  foreach $hairpin (@hairpins)
    {
    $hairpin =~ tr/ACGU/acgt/;
    $c_hairpin = $hairpin;
    $c_hairpin =~ tr/acgt/tgca/;
    push(@c_hairpins,$c_hairpin);
#Reverso
    @bases = split("",$hairpin);
    @reverse = reverse @bases;
    $r_hairpin = join("",@reverse);
    push(@r_hairpins,$r_hairpin);
#Complementar reverso
```

```
      $rc_hairpin = $r_hairpin;
      $rc_hairpin =~ tr/acgt/tgca/;
      push(@rc_hairpins,$rc_hairpin);

      if ($largest_hairpin < length($hairpin))
        {
        $largest_hairpin=length($hairpin);
        }
    }
  }

read_hairpins("zcat $microrna|");
print "Number of hairpins: $hairpin_count \n";
print "Largest hairpin: $largest_hairpin \n";


open(GENOME,"zcat $genome|");

$fasta_comment=<GENOME>;

chomp($line=<GENOME>);
$length_of_line=length($line);
$minimum_lines_of_overlap=$largest_hairpin / $length_of_line;
$minimum_lines_of_overlap=ceil($minimum_lines_of_overlap)+1;
print "Minimum lines of overlap between contigs $minimum_lines_of_overlap \n";

@read_buffer;
$line =~ tr/ACGT/acgt/;
$read_buffer[0]=$line;


for($bf=1; $bf < $lines_in_buffer; $bf++)
  {
  chomp($line=<GENOME>);
  if ($line =~ />.*/)
    {$read_buffer[$bf]="\$";
    }
  else
    {
    $line =~ tr/ACGT/acgt/;
    $read_buffer[$bf]=$line;
    }
  }

$|=1;
$start=0;
$found=0;
while (!eof(GENOME))
  {
  $contig = join("", @read_buffer);
  #print $contig ,"\n\n";
  $count=0; $type=0;
  @hairpin_type=("main","main reverse","complementary","complementary reverse");
  foreach $hairpin (@hairpins,@r_hairpins,@c_hairpins,@rc_hairpins)
    {
    #print $hairpin, "\n\n";
    if ($contig =~ /$hairpin/)
      {
      $found=$found+1;
      $p=$-[0];
      print VIZINHO  "> $found ", substr($hairpins_id[$count],1,150)," at ",$start
      + $p," ",$hairpin_type[$type],"\n";
      print ANTES    "> $found ", substr($hairpins_id[$count],1,150)," at ",$start
      + $p," ",$hairpin_type[$type],"\n";
      print DEPOIS   "> $found ", substr($hairpins_id[$count],1,150)," at ",$start
      + $p," ",$hairpin_type[$type],"\n";
      if (($p-$limite) < 0) #Ops, posicao negativa!
        {
        $pos=0; #Comeca do inicio de tudo
```

```
          $lim=$p; #Le apenas ate onde achou o hairpin
          }
        else {$pos=$p-$limite; $lim=$limite}
        $antes=substr($contig,$pos,$lim);
        $antes =~ tr/acgt/ACGU/;
        $durante=substr($contig,$p,length($hairpin));
        $durante =~ tr/t/u/;
        $depois=substr($contig,$p+length($hairpin),$limite);
        $depois =~ tr/acgt/ACGU/;
        print VIZINHO $antes,$durante,$depois,"\n";
        print ANTES $antes,"\n";
        print DEPOIS $depois,"\n";
        print "!";
        }
      $count++;
      if($count >= scalar(@hairpins_id)) {$count=0; $type=$type+1;}
      }
    print ".";
    for($bf=0; $bf < ($lines_in_buffer-$minimum_lines_of_overlap); $bf++)
      {
      chomp($line=<GENOME>);
      if (!eof(GENOME) && !($line =~ />.*/))
        {
        $start=$start+length($read_buffer[0]);
        shift(@read_buffer);
        $line =~ tr/ACGT/acgt/;
        push(@read_buffer,$line);
        }
      }
  }
```

B.6 razaocg.pl

Este programa contabiliza os nucleotídeos C e G das sequências dos miRNAs/mirtrons e da vizinhança extraída para calcular a razão CG.

```
open(RESULT,">contagemCGlet.dat");
#open(SOMENOS,">menos.dat");
sub read_hairpins{
  open(HAIRPIN,$_[0]) or die 'Nao pude abrir o arquivo';
  my $linha="";
  $hairpin_count=-1;
  while (chomp($linha=<HAIRPIN>))
    {
    if ($linha =~ />.*/)
      {
      $hairpin_count++;
      $hairpins[$hairpin_count]="";
      }
    else
      {
      $hairpins[$hairpin_count]="$hairpins[$hairpin_count]$linha";
      }
    }
  }
read_hairpins("let.dat");
#read_hairpins("46pmir.dat");
#read_hairpins("5tailedmirtrons-150.dat");
#read_hairpins("../mirtron/mirtronsanalisados.dat");
#read_hairpins("resultados.dat");
#read_hairpins ("antes-mirtrons.dat");???PQ NAO RODA?
#read_hairpins("zcat /share/bioinf/database/hairpin-dme.fa.gz|");
#read_hairpins("novosmiRnas.dat");
#read_hairpins("antes-novosmiRnas.dat");
#read_hairpins("depois-novosmiRnas.dat");

#Determinando a frequencia de nucleotideos(C/G) e o comprimento dos hairpins
@hairpins;

@result;
$mais=0;
$menos=0;

foreach $gene (@hairpins)
  {
  @g = split(//, $gene);

  %nuc=("A",0,"C",0,"G",0,"U",0,"a",0,"c",0,"g",0,"u",0);
  foreach $n (@g) {$nuc{$n}++;}
  $cgvizinhos=($nuc{'C'}+$nuc{'G'})/($nuc{'A'}+$nuc{'C'}+$nuc{'G'}+$nuc{'U'});
  $cghairpin=($nuc{'c'}+$nuc{'g'})/($nuc{'a'}+$nuc{'c'}+$nuc{'g'}+$nuc{'u'});

  push(@result,$cghairpin/$cgvizinhos);
  print RESULT '>',' ', $cghairpin/$cgvizinhos,"\n";
  if ($cghairpin/$cgvizinhos > 1) {$mais++;}
  else {$menos++;
#print SOMENOS '>',"\n",$gene,"\n";}
  }
}
print "+$mais, -$menos,\n";

$media=0;
foreach $cg (@result)
  {
  $media=$media+$cg;
  }
```

```
  $media=$media/scalar(@result);

    print  "contagem de CG:",$media,;

  $desvio=0;
  foreach $cg (@result)
    {
    $desvio=$desvio+($cg-$media)**2;
    }

  $desvio=sqrt($desvio/(scalar(@result)-1));

  print "+/-",$desvio,"\n";
```

B.7 rnafold.pl

Este programa calcula a energia livre de Gibbs através do RNAfold.

```
1
  open(SAIDA,">Energialivredps.dat");

  sub hairpin{
    open(HAIRPIN,$_[0]);
6   my $linha="";
    $cont=0;
    $hairpin="";
    $id="";
    while ($linha=<HAIRPIN>)
11    {
       chomp($linha);#remove o enter das linhas do arquivo 15mirtrons.dat
    #if ($linha =~ /([\D]{3}\-mir\-[^\s]*)/)# pega expressoes como dme-mir-1004,
                                            #cel-mir-2a-b, etc.
    #   if ($linha =~/(MI[0-9]{7})/)
16  #if ($linha =~/(cel)/)
     if ($linha=~/(>)/)
        {
         $newid=$1;#$1 armazena as informacoes contidas dentro do prim.
                   #parenteses da linha de cima
21       $cont++;
         if ($hairpin)
           {
           $RNAfold=`echo $hairpin | RNAfold -noPS`;
           $RNAfold =~ /(-[0-9]+\.[0-9]+)/;
26         $energia= $1;
            $t=length($hairpin); #colocado dpois
           $d=$energia/$t;#colocado dpois
           print SAIDA ">$id $d \n";#colocado dpois

31         #print SAIDA ">$id $energia \n";#comentado depois
           }
         $hairpin="";
         $id=$newid;
         }
36
      else
        {
        $hairpin="$hairpin$linha";
         }
41    }
    if ($hairpin)
           {
           $RNAfold=`echo $hairpin | RNAfold -noPS`;
           $RNAfold =~ /(-[0-9]+\.[0-9]+)/;
46         $energia= $1;
           $t=length($hairpin); #colocado dpois
           $d=$energia/$t;#colocado dpois
           print SAIDA ">$id $d \n";#colocado dpois

51         #print SAIDA ">$id $energia \n";#comentado depois


           }
    }
56 hairpin ("dps211.dat");
  #hairpin ("15mirtronviz.dat");
  #hairpin ("13hmirtrons.dat");
  #hairpin ("4mircel.dat")
  #hairpin ("153hairpin.dat")
61 # como funciona o programa: le a 1 linha e executra so faz o primeiro if, ou seja,
  # ele pega o id # le a 2 linha e so faz o else, ou seja, captura a sequencia
  # le a 3 linha, armazena o id e faz o segundo if, ou seja, agora $hairpin
```

```
  # n e mais vazio pois a sequencia foi capturada na etapa anterior. Roda o
  # RNAfold e imprime na saida o id e energia.
66 # linha 33 em diante: este if fora do while e para rodar o RNAfold para a ultima
  # sequencia pois o programa so roda o rnafold qdo encontra um novo id abaixo e
  # dpois do mir-1017 ele n encontra outro id
```

B.8 rnafold2.pl

Este programa calcula a média e o desvio padrão dos valores de energia livre de Gibbs dos miRNA/mirtrons.

```
open(ARQ,'Energialivredps.dat')or die "Nao pode abrir o arquivo";
open(SAIDA,">gibbs.dat");

@seqs=<ARQ>;
chomp(@seqs);
close(ARQ);

$contador=0;
$media=0;
$desvio=0;

foreach $linha (@seqs)
 {
    if($linha =~ />/)
      {
        $p = 0;
        }
     @campos=split(' ',$linha);
     $contador++;
     print SAIDA " >id=$campos[$p] energia=$campos[$p+1] \n";

      }
  close(SAIDA);

print scalar (@seqs),"\n";

open(SAIDA,'gibbs.dat')or die "Nao pode abrir o arquivo";

while ($linha=<SAIDA>)
 {
    chomp($linha);
    if ($linha =~ /(-[0-9]+\.[0-9]+)/)
     { $e=$1;
       $media=$media+$e;
       #print $media,"\n";
             }
     }
$media=$media/scalar(@seqs);
print "media de energia livre:",$media,"\n";

open(SAIDA,'gibbs.dat')or die "Nao pode abrir o arquivo";

while ($linha=<SAIDA>)
 {
    chomp($linha);
    if ($linha =~ /(-[0-9]+\.[0-9]+)/)
     { $e=$1;
      $desvio=$desvio+($e-$media)**2;
      # print $desvio,"\n";
             }
     }
$desvio=sqrt($desvio/scalar(@seqs));
print "+/-",$desvio,"\n";
```

C Artigo publicado na TCBB

Reprodução do artigo Ref. 74.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Comment on "SCS: Signal, Context, and Structure Features for Genome-Wide Human Promoter Recognition"

Denise Fagundes-Lima and Gerald Weber

Abstract—We comment on the flexibility profiles calculated by Zeng et al., and show that these profiles do not represent the local flexibility of the DNA molecule. If one takes into account the physics of elasticity, the averaged flexibility profile show an additional peak which is missed in the original calculation. We show that it is not possible to calculate the flexibility of a 6-mer using tetranucleotide elastic constants, the shortest sequence is a 7-mer. For 6-mers, dinucleotide or trinucleotide parameters are needed. We present calculations for dinucleotide flexibility parameters and show that the same additional peak is present for both 7-mers and 6-mers.

Index Terms—Promoter analysis, DNA elasticity, flexibility profiles

Zeng et al., in their work on promoter recognition [1], introduce an index which uses tetranucleotide flexibility parameters obtained from Ref. 2. These flexibility parameter have the physical dimension of an elastic constant per mol, see Eq. 1 in Ref. 2. As described in their Eqs. (3–5), Zeng et al. use a simple consecutive summation of these elastic constants. Specifically, they use three consecutive tetranucleotide parameters to calculate a 6-mer index f

$$f_i = t_{i,i+3} + t_{i+1,i+4} + t_{i+2,i+5}, \tag{1}$$

where i is the starting position of the 6-mer in the genomic sequence.

To understand how to obtain the elastic constant of a longer sequence from shorter segments, one has to imagine a DNA molecule as composed of a certain number of coupled springs. Since these springs are chained tail-to-head to each other, it is possible to calculate an equivalent elastic constant which represents the longer molecule. For a 7-mer we would use two tetramers with one overlapping position which is where the two tetramers are chained together. It is therefore not possible to obtain a 6-mer elastic constant from tetramers. Furthermore, it the equivalent elastic constant of a 7-mer is obtained from the inverse summation of the elastic constants of the two tetramers. A direct summation as proposed in Eq. (1) would be representative of 3 tetramers in parallel which does not represent the physical configuration of DNA elasticity. As a result, the index f in Eq. (1) does not represent the elastic property of an 6-mer but has become essentially an arbitrary index.

Nevertheless, the index f is built from elastic constants. It is therefore legitimate to ask to which extent does this index compare to a flexibility profile? To answer this question we calculated the equivalent elastic constants for 7-mers using the same human promoter sequences from DBTSS (version 5.2.0) [3] and tetranucleotide parameters from Ref. 1.

For a 7-mer the resulting equivalent elastic constant $t_{1,7}$ is calculated from tetranucleotide elastic constants $t_{i,i+3}$,

$$\frac{1}{t_{1,7}} = \frac{1}{t_{1,4}} + \frac{1}{t_{4,7}}, \tag{2}$$

where the subscripts i, j represent the start and end position of the segment. Eq. 2 results from a straightforward application of Hookes law and is easy to understand intuitively. Given two coupled springs, say one soft and the other rigid, the force exerted will deform much more the softer spring. Overall, the chained springs are easier to deform than each individually. Therefore, the resulting equivalent elastic constant is dominated by the softer part of the elastic constant. Using the same example as in Eqs. (4–5) of Ref. 1, the equivalent elastic constant of a 7-mer becomes

$$\frac{1}{t_{\mathrm{TATAAAA}}} = \frac{1}{t_{\mathrm{TATA}}} + \frac{1}{t_{\mathrm{AAAA}}}. \tag{3}$$

This highlights yet another problematic aspect of the parameter f which is that the central part of the molecule becomes over-represented.

Unavoidably, using either Eq. (1) or Eq. (2) will result in different profiles. In Fig. 1a we show the profiles recalculated from Ref. 1. Using Eq. (2) we observe additional peak in Fig. 1b, close to transcription starting position, which is missed altogether when using Eq. (1). Interestingly, the region around -28 retains its rigid character even when the flexibility is calculated according

- *D. Fagundes-Lima is with the Department of Biological Sciences, Federal University of Ouro Preto, Ouro Preto-MG, Brazil, defalima@gmail.com*
- *G. Weber is with the Department of Physics, Federal University of Minas Gerais, Belo Horizonte-MG, Brazil.*

Digital Object Indentifier 10.1109/TCBB.2011.130

to Eq. (2). This result is surprising and non-trivial since Eq. (2) generally favours softer elastic constants. It also confirms the interpretation from Zeng *et al.* that this region is notably rigid when tetranucleotide parameters from Ref. 2 are used.

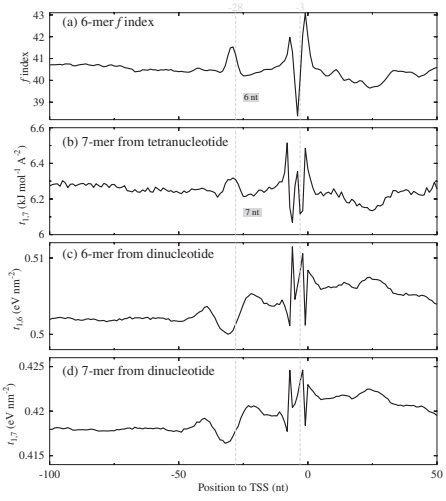


Fig. 1. Promoter profiles for the (a) 6-mer f-index, (b) 7-mer equivalent elastic constant, and elastic constants calculated using dinucleotide parameters for (c) 6-mers and (d) 7-mer. The gray boxes show the width of a 6-mer and a 7-mer sliding window.

The comparison of Figs. 1a and 1b raises two further questions. The first question is whether the additional peak in Fig. 1b does result from Eq. (2) or from a longer sliding window? The next question is how should one proceed to calculate profiles for 6-mers? Starting with the second question, one way to obtain the elastic properties of 6-mers would be to use nucleotide parameters of dimers or trimers. For instance, using dinucleotide flexibility parameters one can generalise Eq. (2) for N-mers

$$\frac{1}{t_{1,N}} = \sum_{i=1,\ldots,N-1} \frac{1}{t_{i,i+1}}. \qquad (4)$$

Figure 1c and 1d shows the flexibility profiles calculated using dinucleotide elastic constants from Ref. 4. Both 6-mer and 7-mer flexibility profile show nearly identical results. Therefore, it seems reasonable to assume that the missed peak of Fig. 1a results from the way the f index is calculated and not from a shorter sliding window.

We take this opportunity to comment on the use of flexibility parameters from different methods and experiments [2], [4], as these provide some interesting insights into the flexibility properties of promoters. Packer *et al.* [2] used data from X-ray diffraction measurements to obtain their elastic constants. This type of measurements probes essentially the static configuration of the DNA molecule and from Fig. 1b one is lead to conclude that the region around -28 should be largely rigid. On the other hand, the parameters from Ref. 4 result from melting temperatures and as such probe the dynamics of the DNA molecule. In this case Figs. 1c,d indicate an exceptionally soft region around -28, in contrast to Fig. 1b.

In conclusion, we show that the f-index is not the appropriate representation of the elasticity of the DNA molecule. However, this does not imply in its inadequacy for promoter recognition. Zeng et al. [1] showed that the f-index is useful for promoter recognition and this remains unchanged. Our finding only concerns the interpretation of these results in terms of DNA elasticity, and draws attention to the problems which arise when comparing f-profiles with flexibility profiles.

ACKNOWLEDGEMENTS

This work was funded by CNPq, Fapemig and National Institute of Science and Technology for Complex Systems.

REFERENCES

[1] J. Zeng, X.-Y. Zhao, X.-Q. Cao, and H. Yan, "SCS: Signal, context, and structure features for genome-wide human promoter recognition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, pp. 550–562, 2010.
[2] M. J. Packer, M. P. Dauncey, and C. A. Hunter, "Sequence-dependent DNA structure: tetranucleotide conformational maps," *J. Mol. Biol.*, vol. 295, no. 1, pp. 85–103, 2000.
[3] R. Yamashita, Y. Suzuki, H. Wakaguri, K. Tsuritani, K. Nakai, and S. Sugano, "DBTSS: database of human transcription start sites, progress report 2006," *Nucleic Acids Research*, vol. 34, no. suppl 1, p. D86, 2006.
[4] G. Weber, J. W. Essex, and C. Neylon, "Probing the microscopic flexibility of DNA from melting temperatures," *Nature Physics*, vol. 5, pp. 769–773, 2009.

Denise Fagundes Lima received the BSc degree in biological sciences from the Federal University of Ouro Preto, Brazil, in 2009. She is now completing her MSc in Biotechnology also at the Federal University of Ouro Preto. Her research interests include bioinformatics and the biology of promoters and microRNA.

Gerald Weber received the PhD degree in physics from the State University of Campinas, Brazil, in 1990. He worked with theoretical semiconductor physics until 2003 when he changed his research interest to the physics of DNA and bioinformatics when at the School of Chemistry of the University of Southampton, UK. Since 2009 he is a lecturer at the Department of Physics of the Federal University of Minas Gerais, Brazil. His current research interest are in thermodynamical properties of DNA and RNA and its application to problems in computational biology.

Referências

[1] Griffiths, A. J. F., Wessler, S. R., Lewontin, R. C., and B.Carroll, S. *Introduction to Genetic Analysis*, volume 9. W H Freeman & Co, (2008).

[2] Gerstein, M., et al. What is a gene, post-ENCODE? history and updated definition. *Genome research* **17**(6), 669–681 (2007).

[3] Pesole, G. What is a gene? An updated operational definition. *Gene* (417), 1–4 (2008).

[4] Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics* **122**(6), 565–581 (2008).

[5] Wilkins, M., Stokes, A., and Wilson, H. Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. *Nature* **171**(4356), 738–740 (1953).

[6] Wray, G., et al. The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution* **20**(9), 1377 (2003).

[7] Jacob, F. and Monod, J. *Genetic regulatory mechanisms in the synthesis of proteins*, 192. Cold Spring Harbor Laboratory Press (1996).

[8] Zhang, M. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics* **3**(9), 698–709 (2002).

[9] Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Research* **30**(1), 328–331 (2002).

[10] Wang, Z., Chen, Y., and Li, Y. A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics* **2**(4), 216–221 (2004).

[11] Zeng, J., Zhao, X.-Y., Cao, X.-Q., and Yan, H. SCS: Signal, context, and structure features for genome-wide human promoter recognition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**, 550–562 (2010).

[12] Pedersen, A., Baldib, P., Chauvinb, Y., and Brunaka, S. The biology of eukaryotic promoter prediction - a review. *Computers & Chemistry* **23**(191), 207 (1999).

[13] Butler, J. and Kadonaga, J. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development* **16**(20), 2583 (2002).

[14] Juven-Gershon, T. and Kadonaga, J. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology* **339**(2), 225–229 (2010).

[15] Geiduschek, E. and Tocchini-Valentini, G. Transcription by RNA polymerase III. *Annual Review of Biochemistry* **57**(1), 873–914 (1988).

[16] Patikoglou, G., et al. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes & Development* **13**(24), 3217 (1999).

[17] Grummt, I. Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. *Genes & Development* **17**(14), 1691 (2003).

[18] Sandelin, A., et al. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics* **8**(6), 424–436 (2007).

[19] Kanhere, A. and Bansal, M. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Research* **33**(10), 3165 (2005).

[20] Fukue, Y., Sumida, N., Nishikawa, J., and Ohyama, T. Core promoter elements of eukaryotic genes have a highly distinctive mechanical property. *Nucleic Acids Research* **32**(19), 5834 (2004).

[21] Roeder, R. The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences* **21**(9), 327–335 (1996).

[22] Smale, S. and Kadonaga, J. The RNA polymerase II core promoter. *Annual Review of Biochemistry* **72**, 449 (2003).

[23] Kutach, A. and Kadonaga, J. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Molecular and Cellular Biology* **20**(13), 4754 (2000).

[24] Baumann, M., Pontiller, J., and Ernst, W. Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: An overview. *Molecular Biotechnology* **45**, 1–7 (2010).

[25] Schug, J., et al. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biology* **6**(4), R33 (2005).

[26] Gershenzon, N., Trifonov, E., and Ioshikhes, I. The features of *Drosophila* core promoters revealed by statistical analysis. *BMC Genomics* **7**(1), 161 (2006).

[27] Juven-Gershon, T., Hsu, J., Theisen, J., and Kadonaga, J. The RNA polymerase II core promoter–the gateway to transcription. *Current Opinion in Cell Biology* **20**(3), 253–259 (2008).

[28] Burke, T. and Kadonaga, J. The downstream core promoter element, DPE, is conserved from drosophila to humans and is recognized by TAFII60 of *Drosophila*. *Genes & Development* **11**(22), 3020 (1997).

[29] Lee, D., et al. Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Molecular and Cellular Biology* **25**(21), 9674 (2005).

[30] Kadonaga, J. The DPE, a core promoter element for transcription by RNA polymerase II. *Experimental and Molecular Medicine* **34**(4), 259–264 (2002).

[31] Deng, W. and Roberts, S. TFIIB and the regulation of transcription by RNA polymerase II. *Chromosoma* **116**(5), 417–429 (2007).

[32] Deng, W. and Roberts, S. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes & Development* **19**(20), 2418 (2005).

[33] Lagrange, T., Kapanidis, A., Tang, H., Reinberg, D., and Ebright, R. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor II B. *Genes & Development* **12**(1), 34 (1998).

[34] Evans, R., Fairley, J., and Roberts, S. Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. *Genes & Development* **15**(22), 2945 (2001).

[35] Wang, Y. and Roberts, S. New insights into the role of TFIIB in transcription initiation. *Transcription* **1**(3), 126 (2010).

[36] Rombauts, S., et al. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiology* **132**(3), 1162 (2003).

[37] Illingworth, R. and Bird, A. CpG islands - A rough guide. *FEBS Letters* **583**(11), 1713–1720 (2009).

[38] Fukue, Y., Sumida, N., Tanase, J., and Ohyama, T. A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance. *Nucleic Acids Research* **33**(12), 3821 (2005).

[39] Yamamoto, Y., Yoshioka, Y., Hyakumachi, M., and Obokata, J. Characteristics of Core Promoter Types with respect to Gene Structure and Expression in *Arabidopsis thaliana*. *DNA Research* **18**(5), 333–342 (2011).

[40] Travers, A. The structural basis of DNA flexibility. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* **362**(1820), 1423 (2004).

[41] Drew, H., Weeks, J., and Travers, A. Negative supercoiling induces spontaneous unwinding of a bacterial promoter. *The EMBO Journal* **4**(4), 1025 (1985).

[42] Breslauer, K. J., Frank, R., Blocker, H., and Marky, L. A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* **83**(11), 3746–3750 (1986).

[43] Harley, C. and Reynolds, R. Analysis of *E. coli* promoter sequences. *Nucleic Acids Research* **15**(5), 2343 (1987).

[44] Kanhere, A. and BansalL, M. DNA bending and curvature: A'turning' point in DNA function? *Proc. Indian Natn. Sci Acad.* **2**, 239–254 (2004).

[45] Dlakić, M. and Harrington, R. The effects of sequence context on DNA curvature. *Proceedings of the National Academy of Sciences* **93**(9), 3847 (1996).

[46] Asayama, M. and Ohyama, T. Curved DNA and prokaryotic promoters. *DNA Conformation and Transcription* , 37–51 (2005).

[47] Venter, J. C., Adams, M. D., Myers, E. W., et al. The Sequence of the Human Genome. *Science* **291**(5507), 1304–1351 (2001).

[48] Pedersen, A., Baldi, P., Chauvin, Y., and Brunak, S. DNA structure in human RNA polymerase II promoters. *Journal of Molecular Biology* **281**(4), 663–673 (1998).

[49] Cao, X., Zeng, J., and Yan, H. Structural properties of replication origins in yeast DNA sequences. *Physical Biology* **5**, 036012 (2008).

[50] Zhang, M., Gong, Y., Osiowy, C., and Minuk, G. Y. Rapid detection of hepatitis B virus mutations using real-time PCR and melting curve analysis. *Hepatology* **36**(3), 723–728 (2002).

[51] Fickett, J. and Hatzigeorgiou, A. Eukaryotic promoter recognition. *Genome Research* **7**(9), 861–878 (1997).

[52] Werner, T. Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome* **10**(2), 168–175 (1999).

[53] Ohler, U. and Niemann, H. Identification and analysis of eukaryotic promoters: recent computational approaches. *TRENDS in Genetics* **17**(2), 56–60 (2001).

[54] Prestridge, D. Predicting Pol II promoter sequences using transcription factor binding sites. *Journal of Molecular Biology* **249**(5), 923–932 (1995).

[55] Down, T. and Hubbard, T. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Research* **12**(3), 458–461 (2002).

[56] Hutchinson, G. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Computer Applications in the Biosciences: CABIOS* **12**(5), 391–398 (1996).

[57] Knudsen, S. Promoter2. 0: for the recognition of Pol II promoter sequences. *Bioinformatics* **15**(5), 356 (1999).

[58] Scherf, M., Klingenhoff, A., and Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *Journal of Molecular Biology* **297**(3), 599–606 (2000).

[59] Brukner, I., Sánchez, R., Suck, D., and Pongor, S. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *The EMBO journal* **14**(8), 1812 (1995).

[60] Satchwell, S., Drew, H., and Travers, A. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology* **191**(4), 659–675 (1986).

[61] El Hassan, M. and Calladine, C. Structural mechanics of bent DNA. *Endeavour* **20**(2), 61–67 (1996).

[62] El Hassan, M. and Calladine, C. Two distinct modes of protein-induced bending in DNA. *Journal of Molecular Biology* **282**(2), 331–343 (1998).

[63] Packer, M. J., Dauncey, M. P., and Hunter, C. A. Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.* **295**(1), 85–103 (2000).

[64] Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., and Van de Peer, Y. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Research* **18**(2), 310–323 (2008).

[65] Florquin, K., Saeys, Y., Degroeve, S., Rouzé, P., and Van de Peer, Y. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Research* **33**(13), 4255–4264 (2005).

[66] Akan, P. and Deloukas, P. DNA sequence and structural properties as predictors of human and mouse promoters. *Gene* **410**(1), 165–176 (2008).

[67] Ohler, U., Liao, G., Niemann, H., Rubin, G., et al. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol* **3**(12), 0081–0087 (2002).

[68] Goñi, J., Pérez, A., Torrents, D., and Orozco, M. Determining promoter location based on DNA structure first-principles calculations. *Genome Biol* **8**(12), R263 (2007).

[69] Wikipedia. Hooke's law. `http://en.wikipedia.org/wiki/Hooke%27s_law#Multiple_springs`.

[70] Weber, G. Sharp DNA denaturation due to solvent interaction. *Europhys. Lett.* **73**(5), 806–811 (2006).

[71] Weber, G., Essex, J. W., and Neylon, C. Probing the microscopic flexibility of DNA from melting temperatures. *Nature Physics* **5**, 769–773 (2009).

[72] Yamashita, R., et al. DBTSS: database of human transcription start sites, progress report 2006. *Nucleic Acids Research* **34**(suppl 1), D86–D89 (2006).

[73] Schmid, C., Praz, V., Delorenzi, M., Périer, R., and Bucher, P. The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Research* **32**(suppl 1), D82–D85 (2004).

[74] Fagundes-Lima, D. and Weber, G. Comment on SCS: Signal, context, and structure features for genome-wide human promoter recognition. *Trans. on Comp. Bio. and Bioinf.* **9**, 940–941 (2012).

[75] Zhang, B., Pan, X., Cobb, G. P., and Anderson, T. A. microRNAs as oncogenes and tumor suppressors. *Developmental Biology* **302**(1), 1–12 (2007).

[76] Lee, R. C., Feinbaum, R. L., and Ambros, V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**(5), 843–854 (1993).

[77] Kim, V., Han, J., and Siomi, M. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology* **10**(2), 126–139 (2009).

[78] Okamura, K., Chung, W. J., and Lai, E. The long and short of inverted repeat genes in animals: microRNAs, mirtrons and hairpin RNAs. *Cell Cycle* **7**(18), 2840 (2008).

[79] Ruby, J., Jan, C., and Bartel, D. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**(7149), 83–86 (2007).

[80] Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., and Lai, E. C. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell* **130**(1), 89–100 (2007).

[81] Berezikov, E., Chung, W., Willis, J., Cuppen, E., and Lai, E. C. Mammalian mirtron genes. *Molecular Cell* **28**(2), 328–336 (2007).

[82] Flynt, A. S., Greimann, J. C., Chung, W. J., Lima, C. D., and Lai, E. C. MicroRNA biogenesis via splicing and exosome-mediated trimming in *Drosophila*. *Molecular Cell* **38**(6), 900–907 (2010).

[83] Martin, R., et al. A drosophila pasha mutant distinguishes the canonical microRNA and mirtron pathways. *Molecular and cellular biology* **29**(3), 861–870 (2009).

[84] Zhang, J., Kuo, C., and Chen, L. GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics* **12**(1), 90 (2011).

[85] Shepard, P. and Hertel, K. Conserved RNA secondary structures promote alternative splicing. *RNA* **14**(8), 1463 (2008).

[86] Weber, G., et al. Thermal equivalence of DNA duplexes without melting temperature calculation. *Nature Physics* **2**, 55–59 (2006).

[87] Morlando, M., et al. Primary microRNA transcripts are processed co-transcriptionally. *Nature Structural & Molecular Biology* **15**(9), 902–909 (2008).

[88] Shomron, N. and Levy, C. MicroRNA-biogenesis and pre-mRNA splicing crosstalk. *Journal of Biomedicine and Biotechnology* **2009** (2009).

[89] Griffiths-Jones, S., Grocock, R., van Dongen, S., Bateman, A., and Enright, A. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140–D144 (2006).

[90] Kozomara, A. and Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research* **39**(suppl 1), D152 (2011).

[91] Hubbard, T., et al. The Ensembl genome database project. *Nucleic Acids Research* **30**(1), 38 (2002).

[92] Drysdale, R. A. and Crosby, M. A. FlyBase: genes and gene models. *Nucleic Acids Research* **33**(suppl 1), D390 (2005).

[93] Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Research* **29**(1), 82 (2001).

[94] Hofacker, I. L. Vienna RNA secondary structure server. *Nucl. Acids. Res.* **31**, 3429–3431 (2003).

[95] Barry Starr, D., Hoopes, B. C., and Hawley, D. K. DNA bending is an important component of site-specific recognition by the TATA binding protein. *J. Mol. Biol.* **250**, 434–446 (1995).

[96] Grove, A., Galeone, A., Mayol, L., and Geiduschek, E. P. Localized DNA flexibility contributes to target site selection by DNA-bending proteins. *J. Mol. Biol.* **260**, 120–125 (1996).